ACL 2020

July 5-10, 2020 / Online

Conference Handbook

UTC+0, Reference City: Accra

The 58th Annual Meeting of the Association for Computational Linguistics

Handbook assembled by Nanyun Violet Peng and Mingyu Derek Ma

Cover designed by Jingya Chen

Contents

Table of Contents **i**

1 Conference Information **1**
Message from the General Chair . . . 1
Message from the Program Committee Co-Chairs . . . 3
Organizing Committee . . . 6
Program Committee . . . 8

2 Tutorials: Sunday, July 5 **11**
Message from the Tutorial Co-Chairs . . . 13
T1: Interpretability and Analysis in Neural NLP . . . 14
T2: Multi-modal Information Extraction from Text, Semi-structured, and Tabular Data on the Web . . . 15
T3: Reviewing Natural Language Processing Research . . . 17
T4: Stylized Text Generation: Approaches and Applications . . . 18
T5: Achieving Common Ground in Multi-modal Dialogue . . . 19
T6: Commonsense Reasoning for Natural Language Processing . . . 20
T7: Integrating Ethics into the NLP Curriculum . . . 21
T8: Open-Domain Question Answering . . . 22

3 Main Conference: Monday, July 6 **23**
Demo Session 1A . . . 26
Session 1A . . . 27
Demo Session 1B . . . 42
Session 1B . . . 43
Demo Session 1C . . . 59
Demo Session 2A . . . 60
Session 2A . . . 61
Demo Session 2B . . . 81
Session 2B . . . 82
Demo Session 2C . . . 100
Demo Session 3A . . . 101

Session 3A . . . 102
Demo Session 3B . . . 120
Session 3B . . . 121
Demo Session 3C . . . 139
Demo Session 4A . . . 140
Keynote Address: Kathleen R. McKeown . . . 141
Session 4A . . . 142
Demo Session 4B . . . 160
Session 4B . . . 161
Demo Session 4C . . . 179
Demo Session 5A . . . 180
Session 5A . . . 181
Demo Session 5B . . . 199
Session 5B . . . 200
Demo Session 5C . . . 216

4 Main Conference: Tuesday, July 7 217
Demo Session 1A . . . 220
Session 6A . . . 221
Demo Session 1B . . . 238
Session 6B . . . 239
Demo Session 1C . . . 255
Demo Session 2A . . . 256
Session 7A . . . 257
Demo Session 2B . . . 277
Session 7B . . . 278
Demo Session 3A . . . 295
Session 8A . . . 296
Demo Session 3B . . . 315
Session 8B . . . 316
Demo Session 3C . . . 336
Demo Session 4A . . . 337
Session 9A . . . 338
Demo Session 4B . . . 357
Session 9B . . . 358
Demo Session 4C . . . 376
Demo Session 5A . . . 377
Session 10A . . . 378
Demo Session 5B . . . 396
Session 10B . . . 397
Demo Session 5C . . . 414

5 Main Conference: Wednesday, July 8 415
Demo Session 1A . . . 418
Session 11A . . . 419
Demo Session 1B . . . 436
Session 11B . . . 437
Demo Session 1C . . . 454
Demo Session 2A . . . 455
Session 12A . . . 456
Demo Session 2B . . . 475
Session 12B . . . 476
Demo Session 2C . . . 494
Demo Session 3A . . . 495
Session 13A . . . 496

Demo Session 3B . . . 515
Session 13B . . . 516
Demo Session 3C . . . 534
Demo Session 4A . . . 535
Keynote Address: Josh Tenenbaum . . . 536
Session 14A . . . 537
Demo Session 4B . . . 555
Session 14B . . . 556
Demo Session 4C . . . 575
Demo Session 5A . . . 576
Session 15A . . . 577
Demo Session 5B . . . 593
Session 15B . . . 594
Demo Session 5C . . . 611

6 Workshops: Sunday-Friday, July 5-July 10 613
WiNLP: The Fourth Widening NLP Workshop . . . 616
IWSLT: The 17th International Conference on Spoken Language Translation . . . 619
NLP4ConvAI: NLP for Conversational AI workshop . . . 621
BioNLP 2020: Workshop on Biomedical Natural Language Processing . . . 622
FEVER: The Third workshop on Fact Extraction and VERification . . . 624
IWPT: The 16th International Conference on Parsing Technologies . . . 625
FLP: The 2nd Workshop on Figurative Language Processing . . . 627
NUSE: The 1st Joint Workshop on Narrative Understanding, Storylines, and Events . . . 629
ALVR: Workshop on Advances in Language and Vision Research . . . 631
RepL4NLP: The 5th Workshop on Representation Learning for NLP . . . 632
NLI: Natural Language Interfaces: Challenges and Promises . . . 635
WNGT: The 4th Workshop on Neural Generation and Translation . . . 636
BEA: The 15th Workshop on Innovative Use of NLP for Building Educational Applications . . . 638
SIGMORPHON: 17th Workshop on Computational Research in Phonetics, Phonology, and Morphology . . . 639
NLPMC: NLP for Medical Conversations . . . 641
ECNLP: The Third Workshop on e-Commerce and NLP . . . 642
SocialNLP: The Eighth International Workshop on Natural Language Processing for Social Media643
AutoSimTrans: The 1st Workshop on Automatic Simultaneous Translation: challenges, recent advances,and future directions . . . 644
Challenge-HML: The Second Grand-Challenge and Workshop on Human Multimodal Language645

7 Anti-harassment Policy 647

8 ACL Author Guidelines 649

Author Index 651

Sponsorship 680

1 Conference Information

Message from the General Chair

It is my delightful duty as General Chair to sit at my kitchen table here in San Francisco and write these words welcoming you to the 58th Annual Meeting of the Association for Computational Linguistics.

Our conference this year is of course very different than in the past; I'll be attending the conference from my kitchen table as well. This is our first experience of ACL as a virtual conference, a shift due to a great trial to all of us, the COVID-19 virus.

Our hope in designing this year's conference was to draw strength from this tragedy and come together as a community. We wanted the conference to offer a beacon of inclusion, making it much easier for people all over the globe, whatever their resources or backgrounds, to come to share their knowledge and learn from each other, in a safe, welcoming, and exciting environment. And we wanted the conference to offer a message of sustainability, proving that even without the environmental costs of thousands of people flying around the globe, and despite the lack of face-to-face cameraderie that helps bind us together, we could nonetheless send our words and thoughts and around the globe and build something together in another way.

Our challenge was to do so in a few months, and with little prior experience of our own. I am so proud of our program chairs Joyce Chai, Natalie Schluter, and Joel Tetreault, and the entire organizing committee, for rising to the challenge and putting together this wonderful meeting.

We have many people to thank. Joyce, Natalie, and Joel, as is our ACL custom, bore the brunt of the organizational burden, and managed beautifully despite all the simultaneous demands of the whirlwinds of their daily work and home lives. The unflappable and wise Priscilla Rasmussen. The amazing 52-person organizing committee, who all turned on a dime to make the conference work virtually: Local Chairs (Jianfeng Gao, Luke Zettlemoyer), Tutorial Chairs (Agata Savary, Yue Zhang), Workshop Chairs (Milica Gašić, Dilek Hakkani-Tur, Saif M. Mohammad, Ves Stoyanov), Student Research Workshop Chairs (Rotem Dror, Jiangming Liu, Shruti Rijhwani, Yizhong Wang), Faculty Advisors to the Student Research Workshop (Omri Abend, Sujian Li, Zhou Yu), Conference Handbook Chair (Nanyun Peng), Demonstration Chairs (Asli Celikyilmaz, Shawn Wen), Diversity and Inclusion Chairs (Cecilia Ovesdotter Alm, Vinodkumar Prabhakaran), Diversity and Inclusion Sub-Committee Chairs (Academic Inclusion Chairs: Aakanksha Naik, Emily Prud'hommeaux, Alla Rozovskaya; Accessibility Chairs: Sushant Kafle, Masoud Rouhizadeh, Naomi Saphra; Childcare Chairs: Khyathi Chandu, Stephen Mayhew; Financial Access Chairs: Allyson Ettinger, Ryan Georgi, Tirthankar Ghosal; Sociocultural Inclusion Chairs: Shruti Palaskar, Maarten Sap), Local Sponsorship Chairs (Hoifung Poon, Kristina Toutanova), Publication Chairs (Steven Bethard, Ryan Cotterell, Rui Yan), Virtual Infrastruc-

ture Chairs (Hao Fang, Sudha Rao), Virtual Infrastructure Committee (Yi Luan, Hamid Palangi, Lianhui Qin, Yizhe Zhang), Publicity Chairs (Emily M. Bender, Esther Seyffarth), Sustainability Chairs (Ananya Ganesh, Klaus Zechner), Student Volunteer Coordinator (Marjan Ghazvininejad), Website Chairs (Sudha Rao, Yizhe Zhang)

The ACL Executive Committee gave excellent guidance and advice. Extra-special thanks to ACL Officers Nitin Madnani, Matt Post, and David Yarowsky. We drew heavily on the infrastructure pioneered by Sasha Rush and the ICLR organization committee at ICLR 2020, together with lots of advice from the organizers of other virtual conferences and the ACM.

We are, as always, extremely grateful to our sponsors, listed on the previous page.

And finally, thanks to you, the thousands of members of our community who made this conference possible by writing papers, recording talks, reviewing and area chairing the papers, being invited speakers, and perhaps most important, by reading

Dan Jurafsky
ACL 2020 General Chair
July 2020

Message from the Program Committee Co-Chairs

Welcome to the 58th Annual Meeting of the Association for Computational Linguistics! ACL 2020 has a special historical significance as this is a particularly exciting period for our field: our field has grown dramatically, NLP research is now ubiquitous in products, and the barrier to entry to the field has lowered considerably. Finally, ACL 2020 is the first ever virtual conference in the community's history. As the world combats the COVID-19 pandemic we are very grateful for all of your support and contributions which make ACL 2020 exciting and memorable.

ACL 2020 received 3,429 submissions–an all-time record for ACL-related conferences! This number represents more than a two-fold increase in submissions from just two years ago. The submissions were assigned to one of 25 topic tracks. This year, we introduced four new tracks: (1) **Ethics and NLP**. Research to assess the associated ethical assumptions and consequences of our NLP applications is crucial as these NLP applications become more and more pervasive and impactful in our society. (2) **Interpretation and Analysis of Models for NLP**. As the community strives to push performance boundaries, understanding behaviors of state-of-the-art models becomes critical. (3) **Theory and Formalism (Linguistic and Mathematical)**. The creation of this track reflects that theoretical research in NLP belongs at ACL and ensures a group of dedicated reviewers for the fair assessment of theory papers. (4) **Theme: Taking Stock of Where We've Been and Where We're Going**. The last few years have witnessed unprecedented growth since the field began over sixty years ago. This track is designed to invite submissions that can provide insight for the community to assess how much we have accomplished today with respect to the past and where the field should be heading.

To meet the reviewer demands of a growing conference without compromising review quality, we initiated a large-scale reviewer recruiting effort. All authors, except for those who explicitly chose to opt-out due to various reasons, were required to review if called upon. We asked all authors to fill out both a global profile and a local profile form that would allow the review system to best detect conflicts of interest (COIs) and to match submissions to reviewers. We thank the overwhelming support from the community. This effort led to a pool of more than 11K candidate reviewers, from which 2,519 primary reviewers were called upon and participated in the review process. Together with Senior Area Chairs (SACs), Area Chairs (ACs), primary reviewers, and secondary reviewers, we have the largest ever program committee in the history of ACL with 3319 members, marking a 47% increase over ACL 2019 (2,256 members).

In addition, we launched a new pilot mentoring program. It is of central importance for our community to mentor and train our new reviewers in order to keep up with the community's rapid growth, both in terms of submissions and in terms of new members of the community, and in order to maintain review quality. In this mentoring program, we pair Area Chairs with mentees (often a Ph.D. student, or a junior researcher who has just graduated) during the review process. The goal is to provide mentoring to new reviewers. The response was very positive. Over 280 ACs and 290 junior reviewers participated in the program. The results of this pilot will inform ACL on constructing more scalable mentoring efforts in the future.

After the review process, 779 papers were accepted which includes 571 long papers and 208 short papers. The acceptance rate is 22.7% based on 3,429 submissions.[1] As in previous years, the acceptance rate for long papers is higher than that for short papers (25.4% vs. 17.6%). Overall, ACL continues to be a highly competitive conference. From the accepted papers, and based on the nominations from Senior Area Chairs, five award-winning papers were selected by a best paper committee, including one best paper and one best theme paper.

Continuing the tradition, ACL 2020 will also feature 31 papers that were published at *Transactions of the Association for Computational Linguistics* (TACL) and, for the first time in ACL history, 7 papers from the journal of *Computational Linguistics* (CL). Another highlight of our program is the two exciting keynote talks: one by Professor Kathleen McKeown from Columbia University, and the other one by Professor Josh Tenenbaum from MIT.

[1] Removing the 29 desk rejects and 312 withdrawals, the acceptance rate becomes 25.2%

Putting together a program for the virtual conference is a new challenge this year. We are fortunate that we were able to learn a lot from ICLR which had a virtual meeting ahead of us. One main issue was making the program accessible to attendees/authors from different time zones. Inspired by the ICLR model, we structured the program with pre-recorded video presentations and live Q&A sessions for individual papers. We thank the authors for providing us their time-slot preferences in a timely manner. Our plenary sessions include live-streamed keynote talks and Q&As, award ceremonies, and business meetings.

ACL 2020 would not be possible without the support from the community. There are many people we would like to thank for their significant contributions!

- Our awesome 40 Senior Area Chairs who were instrumental in every aspect of the review process. For many of them, the scope of their responsibilities was equivalent to chairing a mini-conference. We could always count on them for their input to final decisions, selection of best papers, and outstanding reviewers.
- The 299 Area Chairs who led paper review discussions, wrote meta-reviews, and mentored junior reviewers.
- Our 2,519 primary reviewers and 458 secondary reviewers who provided valuable feedback to the authors. Special thanks to those who stepped in at the last minute to serve as emergency reviewers.
- Our fantastic Best Paper Committee: Christy Doran (chair), Chris Callison-Burch, Yvette Graham, Julia Hirschberg, Rebecca Hwa, Min Yen Kan, Emily Pitler, Dragomir Radev, Philip Resnik, and Yulia Tsvetkov for selecting five award-winning papers under a tight schedule.
- ACL Executive Review Committee. In particular, Amanda Stent and Arya McCarthy for making the COI detection software available and Graham Neubig for the automatic reviewer-paper assignment software. These tools were instrumental in assigning papers to reviewers.
- Our student assistants Shane Storks, Sayan Gosh, Tianchun Huang, Sky Wang, and Tianrong Zhang who helped check the compliance of every single submission.
- Our 7,711 authors who submitted their work for review at ACL 2020. Although we were only able to accept a fraction of the submissions, their hard work makes this conference exciting and our community strong.
- TACL editors-in-chief Mark Johnson, Ani Nenkova, and Brian Roark, TACL Editorial Assistant Cindy Robinson, and CL Editor-in-Chief Hwee Tou Ng for coordinating TACL and CL presentations with us.
- The Program co-Chairs of ACL 2019, Anna Korhonen and David Traum; of NAACL 2019, Christy Doran and Thamar Solorio; of EMNLP 2019, Jing Jiang, Vincent Ng, and Xiaojun Wan for generously sharing their experience, documentation, and advice in organizing ACL conferences and for answering our questions, often on short notice.
- Our Publication Chairs, Steven Bethard, Ryan Cotterell, and Rui Yan, for a smooth transition to the production of the final proceedings.
- Matt Post, the ACL Anthology Director, for his always fast response to our questions.
- Our Publicity Chair, Emily Bender, and our Web Chairs, Sudha Rao and Yizhe Zhang, for effectively communicating conference updates and other useful information.
- Infrastructure Chairs, Hao Feng and Sudha Rao, for taking a heavy load of moving our program online; and Hamid Palangi and Lianhui Qin for coordinating presentations with SlideLive.

- Rich Gerber at SoftConf, who was always quick to respond to our emails and resolve any difficulties we encountered with the START system.
- Priscilla Rasmussen for helpful discussion and insight into organizing an ACL at this scale.
- ICLR chairs, especially Alexander Rush, Shakir Mohamed, and Kyunghyun Cho, for sharing with us many invaluable tips for running a virtual conference.
- ACL Executive Committee, especially Hinrich Schütze, the ACL president, and Barbara Di Eugenio, the liaison for conferences to help us sort through policy issues.
- Our students, interns, postdocs, colleagues, and families. Sorry for ignoring you the past year. We're back!
- And last but not least, our General Chair Dan Jurafsky. He has been open-minded and supportive, giving us the flexibility to innovate while providing an invaluable sounding board, and of course, successfully led the massive turn-around of ACL as a physical conference into a virtual one in just a few short months.

Our deepest gratitude to all of you. We hope you will enjoy this new conference experience.

Joyce Chai, University of Michigan
Natalie Schluter, Google Brain and IT University of Copenhagen
Joel Tetreault, Dataminr

ACL 2020 Program Committee Co-Chairs

Organizing Committee

General Chair
Dan Jurafsky, Stanford University

Program Chairs
Joyce Chai, University of Michigan
Natalie Schluter, Google Brain and IT University of Copenhagen
Joel Tetreault, Dataminr

Local Chairs
Jianfeng Gao, Microsoft Research
Priscilla Rasmussen, ACL
Luke Zettlemoyer, University of Washington

Tutorial Chairs
Agata Savary, University of Tours
Yue Zhang, Westlake University

Workshop Chairs
Milica Gašić, Heinrich Heine University of Düsseldorf
Dilek Hakkani-Tur, Amazon Alexa AI
Saif M. Mohammad, National Research Council Canada
Ves Stoyanov, Facebook AI

Student Research Workshop Chairs
Rotem Dror, Technion - Israel Institute of Technology
Jiangming Liu, The University of Edinburgh
Shruti Rijhwani, Carnegie Mellon University
Yizhong Wang, University of Washington

Faculty Advisors to the Student Research Workshop
Omri Abend, Hebrew University of Jerusalem
Sujian Li, Peking University
Zhou Yu, University of California, Davis

Conference Handbook Chair
Nanyun Violet Peng, UCLA

Demonstration Chairs
Asli Celikyilmaz, Microsoft Research, Redmond
Shawn Wen, PolyAI

Virtual Infrastructure Chairs
Hao Fang, Microsoft Semantic Machines
Sudha Rao, Microsoft Research, Redmond

Virtual Infrastructure Committee
Yi Luan, Google AI Language
Hamid Palangi, Microsoft Research, Redmond
Lianhui Qin, University of Washington
Yizhe Zhang, Microsoft Research, Redmond

Diversity & Inclusion (D&I) Chairs
Cecilia Ovesdotter Alm, Rochester Institute of Technology
Vinodkumar Prabhakaran, Google

Local Sponsorship Chairs
Hoifung Poon, Microsoft
Kristina Toutanova, Google

Publication Chairs
Steven Bethard, University of Arizona
Ryan Cotterell, University of Cambridge

Rui Yan, Peking University

Publicity Chairs
Emily M. Bender, University of Washington
Esther Seyffarth, University of Düsseldorf
Zhiyuan Liu, Tsinghua University

Sustainability Chair
Ananya Ganesh, Educational Testing Service
Klaus Zechner, Educational Testing Service

Student Volunteer Coordinator
Marjan Ghazvininejad, Facebook AI Lab

Website & Conference App Chairs
Sudha Rao, Microsoft Research, Redmond
Yizhe Zhang, Microsoft Research, Redmond

Best Paper Committee
Christy Doran (chair), Clockwork Language
Chris Callison-Burch, University of Pennsylvania
Yvette Graham, Dublin City University
Julia Hirschberg, Columbia University
Rebecca Hwa, University of Pittsburgh
Min Yen Kan, National University of Singapore
Emily Pitler, Google Research
Dragomir Radev, Yale University
Phil Resnik, University of Maryland
Yulia Tsvetkov, Carnegie Mellon University

Program Committee

Program Chairs
Joyce Chai, University of Michigan
Natalie Schluter, Google Brain and IT University of Copenhagen
Joel Tetreault, Dataminr

Senior Area Chairs and Area Chairs
(Senior area chairs are in bold.)

Cognitive Modeling and Psycholinguistics
Emily Prud'hommeaux, Cassandra L. Jacobs, Cecilia Ovesdotter Alm, Christos Christodoulopoulos, Masoud Rouhizadeh, Serguei Pakhomov, Yevgeni Berzak

Computational Social Science and Social Media
Tim Baldwin, **Nikolaos Aletras**, A. Seza Dögruöz, Afshin Rahimi, Alice Oh, Brendan O'Connor, Daniel Preotiuc-Pietro, David Bamman, David Jurgens, David Mimno, Diana Inkpen, Diyi Yang, Eiji Aramaki, Jacob Eisenstein, Jonathan K. Kummerfeld, Kalina Bontcheva

Dialogue and Interactive Systems
Jason Williams, **Mari Ostendorf**, Alborz Geramifard, Amanda Stent, Asli Celikyilmaz, Casey Kennington, David Traum, Dilek Hakkani-Tur, Gabriel Skantze, Helen Hastie, Heriberto Cuayahuitl, Kai Yu, Kallirroi Georgila, Luciana Benotti, Luis Fernando D'Haro, Nina Dethlefs, Ryuichiro Higashinaka, Stefan Ultes, Sungjin Lee, Tsung-Hsien Wen, Y-Lan Boureau, Yun-Nung Chen, Zhou Yu

Discourse and Pragmatics
Annie Louis (taking over for **Diane Litman**), Chloé Braud, Junyi Jessy Li, Manfred Stede, Shafiq Joty, Sujian Li, Yangfeng Ji

Ethics and NLP
Dirk Hovy, Alan W Black, Emily M. Bender, Vinodkumar Prabhakaran, Yulia Tsvetkov

Generation
Wei Xu, **Alexander Rush**, John Wieting, Laura Perez-Beltrachini, Lu Wang, Miltiadis Allamanis, Mohit Iyyer, Nanyun Peng, Sam Wiseman, Shashi Narayan, Sudha Rao, Tatsunori Hashimoto, Xiaojun Wan, Xipeng Qiu

Information Extraction
Doug Downey, **Hoifun Poon**, Alan Ritter, Chandra Bhagavatula, Gerard de Melo, Kai-Wei Chang, Marius Pasca, Mo Yu, Radu Florian, Ruihong Huang, Sameer Singh, Satoshi Sekine, Snigdha Chaturvedi, Sumithra Velupillai, Timothy Miller, Vivek Srikumar, William Yang Wang, Yunyao Li

Information Retrieval and Text Mining
Chin-Yew Lin, **Nazli Goharian**, Andrew Yates, Arman Cohan, Bing Qin, Craig Macdonald, Danai Koutra, Elad Yom-Tov, Franco Maria Nardini, Kalliopi Zervanou, Luca Soldaini, Nicola Tonellotto, Pu-Jen Cheng, Seung-won Hwang, Yangqiu Song, Yansong Feng

Interpretability and Analysis of Models for NLP
Yoav Goldberg, Adina Williams, Afra Alishahi, Douwe Kiela, Grzegorz Chrupała, Marco Baroni, Yonatan Belinkov, Zachary C. Lipton

Language Grounding to Vision, Robotics and Beyond
Yoav Artzi, Angeliki Lazaridou, Dan Goldwasser, Jason Baldridge, Jesse Thomason, Lisa Anne Hendricks, Parisa Kordjamshidi, Raffaella Bernardi, Vicente Ordonez, Yonatan Bisk

Machine Learning for NLP
André Martins, **Isabelle Augenstein**, Ankur Parikh, Anna Rumshisky, Bruno Martins, Caio Corro, Dani Yogatama, Daniel Beck, Dipanjan Das, Edouard Grave, Emma Strubell, Gholamreza Haffari, Ivan Titov, Joseph Le Roux, Jun Suzuki, Kevin Gimpel, Michael Auli, Ming-Wei Chang, Shay B. Cohen, Vlad Niculae, Waleed Ammar, Wilker Aziz, Yejin Choi, Zita Marinho, Zornitsa Kozareva

Machine Translation
Marine Carpuat, **Alexandra Birch**, Ann Clifton, Antonio Toral, Atsushi Fujita, Boxing Chen, Carolina Scarton, Chi-kiu Lo, Christian Hardmeier, Deyi Xiong, Franois Yvon, George Foster, Jiajun

Zhang, Jrg Tiedemann, Maja Popovič, Marcello Federico, Marcin Junczys-Dowmunt, Marco Turchi, Marta R. Costa-jussà, Matt Post, Nadir Durrani, Qun Liu, Rico Sennrich, Taro Watanabe, Yuki Arase, Yvette Graham

Multidisciplinary and Area Chair COI

Michael Strube, Anders Søgaard, David Schlangen, Katrin Erk, Kentaro Inui, Kevin Duh, Massimo Poesio, Mausam, Pascal Denis

NLP Applications

Preslav Nakov, **Karin Verspoor**, Alexander Fraser, Antonio Jimeno Yepes, Aoife Cahill, Daniel Cer, Diarmuid Ó Séaghdha, Giovanni Da San Martino, Hassan Sajjad, Kevin Cohen, Marcos Zampieri, Michel Galley, Min Zhang, Pierre Zweigenbaum, Razvan Bunescu, Sara Rosenthal, Tristan Naumann, Vincent Ng, Wei Gao, Wei Lu

Phonology, Morphology and Word Segmentation

Kemal Oflazer, Christo Kirov, David R. Mortensen, Kareem Darwish, Reut Tsarfaty, Yue Zhang, Özlem Çetinoğlu

Question Answering

Eugene Agichtein, **Alessandro Moschitti**, Avi Sil, Dina Demner-Fushman, Evangelos Kanoulas, Gerhard Weikum, Idan Szpektor, Jimmy Lin, Oleg Rokhlenko, Sanda Harabagiu, Wen-tau Yih, William Cohen

Resources and Evaluation

Nathan Schneider, **Barbara Plank**, Allyson Ettinger, Annemarie Friedrich, Antonios Anastasopoulos, Arianna Bisazza, Claire Bonial, Daniel Zeman, Emmanuele Chersoni, Ines Rehbein, Lonneke van der Plas, Maria Liakata, Sara Tonelli, Sarvnaz Karimi, Tim Van de Cruys, Vered Shwartz, Walid Magdy, Çağri Çöltekin

Semantics: Lexical

Ekaterina Shutova, **Aline Villavicencio**, Alessandro Lenci, Anna Feldman, Aurélie Herbelot, Beata Beigman Klebanov, Carlos Ramisch, Chris Biemann, Enrico Santus, Fabio Massimo Zanzotto, Helen Yannakoudakis, Ivan Vulič, Jose Camacho-Collados, Marianna Apidianaki, Paul Cook, Saif Mohammad

Semantics: Sentence Level

Mohit Bansal, Andreas Vlachos, Christopher Potts, Danqi Chen, Eunsol Choi, He He, Jonathan Berant, Kevin Small, Marek Rei, Sebastian Ruder, Siva Reddy, Swabha Swayamdipta, Thomas Wolf, Veselin Stoyanov

Semantics: Textual Inference and Other Areas of Semantics

Sam Bowman, Anette Frank, Eduardo Blanco, Edward Grefenstette, Jacob Andreas, Jonathan May, Kenton Lee, Lasha Abzianidze, Luheng He, Mehrnoosh Sadrzadeh, Rachel Rudinger, Roy Schwartz, Valeria de Paiva

Sentiment Analysis, Stylistic Analysis, and Argument Mining

Smaranda Muresan, **Swapna Somasundaran**, Bing Liu, Claire Cardie, Elena Musi, Iryna Gurevych, Julian Brooke, Lun-Wei Ku, Marie-Francine Moens, Minlie Huang, Paolo Rosso, Roman Klinger, Serena Villata, Soujanya Poria, Thamar Solorio, Yulan He

Speech and Multimodality

Eric Fosler-Lussier, Bhuvana Ramabhadran, Florian Metze, Gerasimos Potamianos, Hamid Palangi, Martha Larson

Summarization

Fei Liu, Caiming Xiong, Giuseppe Carenini, Katja Markert, Manabu Okumura, Michael Elhadad, Ramesh Nallapati, Sebastian Gehrmann, Wenjie Li, Xiaodan Zhu, Yang Gao

Syntax: Tagging, Chunking and Parsing

David Chiang, Carlos Gómez-Rodríguez, Emily Pitler, Liang Huang, Miguel Ballesteros, Miryam de Lhoneux, Slav Petrov, Stephan Oepen, Weiwei Sun

Theme

Marilyn Walker (taking over for **Ellen Riloff**), Donia Scott, Johan Bos, Luke Zettlemoyer, Philipp

Koehn, Raymond Mooney

Theory and Formalism in NLP (Linguistic and Mathematical)

Daniel Gildea, Alexander Koller, Laura Kallmeyer, Marco Kuhlmann

2

Tutorials: Sunday, July 5

Overview

13:00–16:30 **Tutorials Session 1**

Interpretability and Analysis in Neural NLP (Cutting-edge) [Website]
Yonatan Belinkov, Sebastian Gehrmann, and Ellie Pavlick

Achieving Common Ground in Multi-modal Dialogue (Cutting-edge) [Website]
Malihe Alikhani and Matthew Stone

Integrating Ethics into the NLP Curriculum (Introductory) [Website]
Emily M. Bender, Dirk Hovy, and Alexandra Schofield

17:30–21:00 **Tutorials Session 2**

Interpretability and Analysis in Neural NLP (Cutting-edge) [Website]
Yonatan Belinkov, Sebastian Gehrmann, and Ellie Pavlick

Reviewing Natural Language Processing Research (Introductory) [Website]
Kevin Cohen, Karën Fort, Margot Mieskes, and Aurélie Névéol

Stylized Text Generation: Approaches and Applications (Cutting-edge) [Website]
Lili Mou and Olga Vechtomova

Integrating Ethics into the NLP Curriculum (Introductory) [Website]
Emily M. Bender, Dirk Hovy, and Alexandra Schofield

22:00–1:30 (+1) **Tutorials Session 3**

Multi-modal Information Extraction from Text, Semi-structured, and Tabular Data on the Web (Cutting-edge) [Website]
Xin Luna Dong, Hannaneh Hajishirzi, Colin Lockard, and Prashant Shiralkar

Achieving Common Ground in Multi-modal Dialogue (Cutting-edge) [Website]
Malihe Alikhani and Matthew Stone

Commonsense Reasoning for Natural Language Processing (Introductory) [Website]
Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth

Open-Domain Question Answering (Cutting-edge) [Website]
Danqi Chen and Wen-tau Yih

Message from the Tutorial Co-Chairs

Welcome to the Tutorials Session of ACL 2020.

The ACL tutorials session is organized to give conference attendees a comprehensive introduction by expert researchers to some topics of importance drawn from our rapidly growing and changing research field.

This year, as has been the tradition over the past few years, the call, submission, reviewing and selection of tutorials were coordinated jointly for multiple conferences: ACL, AACL-IJCNLP, COLING and EMNLP. We formed a review committee of 19 members, including the ACL tutorial chairs (Agata Savary and Yue Zhang), the EMNLP tutorial chairs (Banjamin van Durme and Aline Villavicencio), the COLING tutorial chairs (Daniel Beck and Lucia Specia), the AACL-IJCNMP tutorial chairs (Timothy Baldwin and Fei Xia) and 11 external reviewers (Emily Bender, Erik Cambria, Gaël Dias, Stefan Evert, Yang Liu, João Sedoc, Xu Sun, Yulia Tsvetkov, Taro Watanabe, Aaron Steven White and Meishan Zhang). A reviewing process was organised so that each proposal receives 3 reviews. The selection criteria included clarity, preparedness, novelty, timeliness, instructors' experience, likely audience, open access to the teaching materials, diversity (multilingualism, gender, age and geolocation) and the compatibility of preferred venues. A total of 43 tutorial submissions were received, of which 8 were selected for presentation at ACL.

We solicited two types of tutorials, including cutting-edge themes and introductory themes. The 8 tutorials for ACL include of 3 introductory tutorials and 5 cutting-edge tutorials. The introductory tutorials are dedicated to reviewing, ethics and commonsense reasoning in NLP. The cutting-edge discussions address interpretability of neural NLP, multi-modal information extraction and dialogue, stylized text generation, and open-domain question answering.

We would like to thank the tutorial authors for their contributions and flexibility while organising the conference virtually. We are also grateful to the 11 external reviewers for their generous help in the decision process. Finally, our thanks go to the conference organizers for effective collaboration, and in particular to the general chair Dan Jurafsky, the website chairs Sudha Rao and Yizhe Zhang, the publicity chair Emily Bender, the ACL anthology director Matt Post.

We hope you enjoy the tutorials.

ACL 2020 Tutorial Co-chairs
Agata Savary
Yue Zhang

Tutorial 1

Interpretability and Analysis in Neural NLP

Cutting-edge

Yonatan Belinkov, Sebastian Gehrmann, and Ellie Pavlick

[Website]

While deep learning has transformed the natural language processing (NLP) field and impacted the larger computational linguistics community, the rise of neural networks is stained by their opaque nature: It is challenging to interpret the inner workings of neural network models, and explicate their behavior. Therefore, in the last few years, an increasingly large body of work has been devoted to the analysis and interpretation of neural network models in NLP. This body of work is so far lacking a common framework and methodology. Moreover, approaching the analysis of modern neural networks can be difficult for newcomers to the field. This tutorial aims to fill this gap and introduce the nascent field of interpretability and analysis of neural networks in NLP. The tutorial will cover the main lines of analysis work, such as structural analyses using probing classifiers, behavioral studies and test suites, and interactive visualizations. We will highlight not only the most commonly applied analysis methods, but also the specific limitations and shortcomings of current approaches, in order to inform participants where to focus future efforts.

Yonatan Belinkov, Postdoctoral Fellow, Harvard University and MIT

email: belinkov@seas.harvard.edu
website: http://people.csail.mit.edu/belinkov
Yonatan Belinkov is a Postdoctoral Fellow at the Harvard School of Engineering and Applied Sciences (SEAS) and the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL). His research interests are in interpretability and robustness of neural models of language. He has done previous work in machine translation, speech recognition, community question answering, and syntactic parsing. His research has been published at ACL, EMNLP, NAACL, CL, TACL, ICLR, and NeurIPS. His PhD dissertation at MIT analyzed internal language representations in deep learning models. He co-organized or co-organizes BlackboxNLP 2019, BlackboxNLP 2020, and the WMT 2019 machine translation robustness task, and serves as an area chair for the analysis and interpretability track at ACL and EMNLP 2020.

Sebastian Gehrmann, Research Scientist, Google AI

email: gehrmann@google.com
website: http://sebastiangehrmann.com
Sebastian is research scientist at Google AI. He received his PhD in 2020 from Harvard University. His research focuses on the development and evaluation of controllable and interpretable models for language generation. By applying methods from human-computer interaction and visualization to problems in NLP, he develops interactive interfaces that help with the interpretation and explanation of neural networks. His research has been published at ACL, NAACL, EMNLP, CHI, and IEEE VIS. He received an honorable mention at VAST 2018 and was nominated for ACL best demo 2019 for his work on interactive visualization tools. He co-organized INLG 2019 and served as an area chair in summarization for ACL 2020.

Ellie Pavlick, Assistant Professor of Computer Science, Brown University

email: ellie_pavlick@brown.edu
website: http://cs.brown.edu/people/epavlick
Ellie Pavlick is an Assistant Professor at Brown University and a Research Scientist at Google. She received her PhD in 2017 with her thesis on modeling compositional lexical semantics. Her current work focuses on computational models of semantics and pragmatics, with a focus on building cognitively-plausible representations. Her recent work has focused on "probing" distributional models in order to better understand the linguistic phenomena that are and are not encoded "for free" via language modelling. Her work has been published at ACL, NAACL, EMNLP, TACL, *SEM, and ICLR, including two best paper awards at *SEM 2016 and 2019. Ellie co-organized the 2018 JSALT summer workshop on building and evaluating general-purpose sentence representations. She also served as area chair for ACL's sentencelevel semantics track.

Tutorial 2

Multi-modal Information Extraction from Text, Semi-structured, and Tabular Data on the Web

Cutting-edge

Xin Luna Dong, Hannaneh Hajishirzi, Colin Lockard, and Prashant Shiralkar

[Website]

The World Wide Web contains vast quantities of textual information in several forms: unstructured text, template-based semi-structured webpages (which present data in key-value pairs and lists), and tables. Methods for extracting information from these sources and converting it to a structured form have been a target of research from the natural language processing (NLP), data mining, and database communities. While these researchers have largely separated extraction from web data into different problems based on the modality of the data, they have faced similar problems such as learning with limited labeled data, defining (or avoiding defining) ontologies, making use of prior knowledge, and scaling solutions to deal with the size of the Web. In this tutorial we take a holistic view toward information extraction, exploring the commonalities in the challenges and solutions developed to address these different forms of text. We will explore the approaches targeted at unstructured text that largely rely on learning syntactic or semantic textual patterns, approaches targeted at semi-structured documents that learn to identify structural patterns in the template, and approaches targeting web tables which rely heavily on entity linking and type information. While these different data modalities have largely been considered separately in the past, recent research has started taking a more inclusive approach toward textual extraction, in which the multiple signals offered by textual, layout, and visual clues are combined into a single extraction model made possible by new deep learning approaches. At the same time, trends within purely textual extraction have shifted toward full-document understanding rather than considering sentences as independent units. With this in mind, it is worth considering the information extraction problem as a whole to motivate solutions that harness textual semantics along with visual and semi-structured layout information. We will discuss these approaches and suggest avenues for future work.

In alphabetical order,

Xin Luna Dong is a Principal Scientist at Amazon, leading the efforts of constructing Amazon Product Knowledge Graph. She was one of the major contributors to the Google Knowledge Vault project, and has led the Knowledge-based Trust project, which is called the "Google Truth Machine" by the Washington Post. She co-authored the book "Big Data Integration", was awarded ACM Distinguished Member, VLDB Early Career Research Contribution Award for "advancing the state of the art of knowledge fusion", and Best Demo award in Sigmod 2005. She serves on the VLDB endowment and PVLDB advisory committee, and was a PC co-chair for VLDB 2021, ICDE Industry 2019, VLDB Tutorial 2019, Sigmod 2018 and WAIM 2015. She has given multiple tutorials on data integration, graph mining, and knowledge management.

Email: lunadong@amazon.com
Homepage: http://lunadong.com/

Hannaneh Hajishirzi is an Assistant Professor at the Paul G. Allen School of Computer Science & Engineering at the University of Washington. She works on NLP, AI, and machine learning, particularly designing algorithms for semantic understanding, reasoning, question answering, and information extraction from multimodal data. She has earned numerous awards for her research, including an Allen Distinguished Investigator Award, a Google Faculty Research Award, a Bloomberg Data Science Award, an Amazon Research Award, and a SIGDIAL Best Paper Award.

Email: hannaneh@washington.edu
Homepage: https://homes.cs.washington.edu/hannaneh/

Colin Lockard is a PhD student at the Paul G. Allen School of Computer Science & Engineering at the University of Washington, where he has published papers on knowledge extraction from both unstructured and semi-structured text.

Email: lockardc@cs.washington.edu
Homepage: https://homes.cs.washington.edu/lockardc/

Prashant Shiralkar is an Applied Scientist in the Product Graph team at Amazon. He currently works on knowledge extraction from semistructured data. Previously, he received a Ph.D. from Indiana University Bloomington where his dissertation work focused on devising computational approaches for fact checking by mining knowledge graphs. His research interests include machine learning, data mining, information extraction and NLP, and Semantic Web technologies.

Email: shiralp@amazon.com
Homepage: https://sites.google.com/site/shiralkarprashant

Tutorial 3

Reviewing Natural Language Processing Research

Introductory

Kevin Cohen, Karën Fort, Margot Mieskes, and Aurélie Névéol

[Website]

This tutorial will cover the theory and practice of reviewing research in natural language processing. Heavy reviewing burdens on natural language processing researchers have made it clear that our community needs to increase the size of our pool of potential reviewers. Simultaneously, notable "false negatives"—rejection by our conferences of work that was later shown to be tremendously important after acceptance by other conferences—have raised awareness of the fact that our reviewing practices leave something to be desired. We do not often talk about "false positives" with respect to conference papers, but leaders in the field have noted that we seem to have a publication bias towards papers that report high performance, with perhaps not much else of interest in them. It need not be this way. Reviewing is a learnable skill, and you will learn it here via lectures and a considerable amount of hands-on practice.

In alphabetical order,

Kevin Bretonnel Cohen has written, overseen, and received hundreds of reviews in his capacity as deputy editor-in-chief of a biomedical informatics journal, associate editor of five natural language processing or bioinformatics journals, special issue editor, workshop organizer, and author of 100+ publications in computational linguistics and natural language processing. His forthcoming book Writing about data science research: With examples from machine and natural language processing includes coverage of a number of aspects of the reviewing process. His current research focuses on issues of reproducibility.

Karën Fort is an associate professor at Sorbonne Universite. Besides being a reviewer for most major NLP conferences, she has been editor in chief for a Traitement automatique des langues journal special issue on ethics and acted as Area Chair for ACL in 2017 and 2018 (as senior AC). Her main research interests are ethics, and the construction of language resources for natural language processing. She co-authored the report on the EMNLP reviewer survey (Névéol et al., 2017).

Margot Mieskes is a professor at the Darmstadt University of Applied Sciences and as such has a lot experience teaching, also in culturally diverse settings, which are prevalent in German Universities of Applied Sciences. Additionally, she has written and received a number of reviews in conferences as well as journals. She is a member of the ACL Professional Conduct Committee and an active member of the Widening NLP efforts. Her research interests are in summarization and summarization evaluation, replicability, repeatability and transparency of NLP experiments in general.

Aurélie Névéol is a permanent researcher at LIMSI CNRS and Universite Paris Saclay. She ´ has been involved in reviewing natural language processing papers at many stages of the reviewing process, including: reviewer, associate editor for three journals, area chair for *ACL and bioinformatics conferences, workshop organizer. Her research focuses on biomedical natural language processing as well as ethics issues in NLP research. She co-authored the report on EMNLP reviewer survey (Névéol et al., 2017).

Tutorial 4

Stylized Text Generation: Approaches and Applications

Cutting-edge

Lili Mou and Olga Vechtomova

[Website]

Text generation has played an important role in various applications of natural language processing (NLP), and kn recent studies, researchers are paying increasing attention to modeling and manipulating the style of the generation text, which we call stylized text generation. In this tutorial, we will provide a comprehensive literature review in this direction. We start from the definition of style and different settings of stylized text generation, illustrated with various applications. Then, we present different settings of stylized generation, such as style-conditioned generation, style-transfer generation, and style-adversarial generation. In each setting, we delve deep into machine learning methods, including embedding learning techniques to represent style, adversarial learning, and reinforcement learning with cycle consistency to match content but to distinguish different styles. We also introduce current approaches to evaluating stylized text generation systems. We conclude our tutorial by presenting the challenges of stylized text generation and discussing future directions, such as small-data training, non-categorical style modeling, and a generalized scope of style transfer (e.g., controlling the syntax as a style).

Lili Mou

doublepower.mou@gmail.com

https://lili-mou.github.io

Dr. Lili Mou is an Assistant Professor at the Department of Computing Science, University of Alberta. He is also an Amii Fellow and a Canadian CIFAR AI Chair. Lili received his BS and PhD degrees in 2012 and 2017, respectively, from School of EECS, Peking University. After that, he worked as a postdoctoral fellow at the University of Waterloo and a research scientist at Adeptmind (a startup in Toronto, Canada). His research interests include deep learning applied to natural language processing as well as programming language processing. Recently, he has been focusing more on text generation, from both continuous latent space and discrete word space. He has more than 30 papers published at top-tier conferences and journals, including AAAI, ACL, CIKM, COLING, EMNLP, ICASSP, ICLR, ICML, IJCAI, INTERSPEECH, NAACL-HLT, and TACL. He presented a tutorial "Discreteness in Neural Natural Language Processing" at EMNLP-IJCNLP'19.

Olga Vechtomova

ovechtomova@uwaterloo.ca

https://ov-research.uwaterloo.ca

Dr. Olga Vechtomova is an Associate Professor in the Department of Management Sciences, Faculty of Engineering, cross-appointed in the School of Computer Science at the University of Waterloo. Olga leads the Natural Language Processing Lab, affiliated with the Waterloo.AI Institute. Her research has been supported by a number of industry and government grants, including Amazon Research Award and Natural Sciences and Engineering Research Council (NSERC). The research in her Lab is mainly focused on designing deep neural networks for natural language generation tasks. Her current and recent projects include controlled text generation, text style transfer, and designing text generative models for creative applications. She has over 50 publications in NLP and Information Retrieval conferences and journals, including NAACL-HLT, COLING, ACL, ACM SIGIR, and CIKM. She and her colleagues recently received the ACM SIGIR 2019 Test of Time Award.

Tutorial 5

Achieving Common Ground in Multi-modal Dialogue

Cutting-edge

Malihe Alikhani and Matthew Stone

[Website]

All communication aims at achieving common ground (grounding): interlocutors can work together effectively only with mutual beliefs about what the state of the world is, about what their goals are, and about how they plan to make their goals a reality. Computational dialogue research offers some classic results on grouding, which unfortunately offer scant guidance to the design of grounding modules and behaviors in cutting-edge systems. In this tutorial, we focus on three main topic areas: 1) grounding in human-human communication; 2) grounding in dialogue systems; and 3) grounding in multi-modal interactive systems, including image-oriented conversations and human-robot interactions. We highlight a number of achievements of recent computational research in coordinating complex content, show how these results lead to rich and challenging opportunities for doing grounding in more flexible and powerful ways, and canvass relevant insights from the literature on human–human conversation. We expect that the tutorial will be of interest to researchers in dialogue systems, computational semantics and cognitive modeling, and hope that it will catalyze research and system building that more directly explores the creative, strategic ways conversational agents might be able to seek and offer evidence about their understanding of their interlocutors.

Malihe Alikhani is a 5th year Ph.D. student in the department of Computer sSience at Rutgers University, advised by Prof. Matthew Stone. She is pursuing a certificate in cognitive science through the Rutgers Center for Cognitive Science and holds a BA and MA in Mathematics. Her research aims at teaching machines to understand and generate multimodal communication. She is the recipient of the fellowship award for excellence in computation and data sciences from Rutgers Discovery Informatics Institute in 2018 and the Anita Berg student fellowship in 2019. Before joining Rutgers, she was a lecturer and an adjunct professor of Mathematics and Statistics for a year at San Diego State University and San Diego Mesa College. She has served as the program committee of ACL, NAACL, EMNLP, AAAI, ICRL, ICMI, and INLG and is currently the associate editor of the Mental Note Journal.

email: `ma1195@cs.rutgers.edu`
webpage: `www.malihealikhani.com`

Matthew Stone is professor and chair in the Department of Computer Science at Rutgers University; he holds a joint appointment in the Rutgers Center for Cognitive Science. His research focuses on discourse, dialogue and natural language generation; he is particularly interested in leveraging semantics to make interactive systems easier to build and more human-like in their behavior. He was program co-chair for NAACL 2007, general co-chair for SIGDIAL 2014. He has also served as program co-chair for INLG and IWCS, as an information officer for SIGSEM, and on the editorial board for Computational Linguistics.

email: `mdstone@cs.rutgers.edu`
website: `www.cs.rutgers.edu/ mdstone/`

Tutorial 6

Commonsense Reasoning for Natural Language Processing

Introductory

Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth

[Website]

Commonsense knowledge, such as knowing that "bumping into people annoys them" or "rain makes the road slippery", helps humans navigate everyday situations seamlessly. Yet, endowing machines with such human-like commonsense reasoning capabilities has remained an elusive goal of artificial intelligence research for decades. In recent years, commonsense knowledge and reasoning have received renewed attention from the natural language processing (NLP) community, yielding exploratory studies in automated commonsense understanding. We organize this tutorial to provide researchers with the critical foundations and recent advances in commonsense representation and reasoning, in the hopes of casting a brighter light on this promising area of future research. In our tutorial, we will (1) outline the various types of commonsense (e.g., physical, social), and (2) discuss techniques to gather and represent commonsense knowledge, while highlighting the challenges specific to this type of knowledge (e.g., reporting bias). We will then (3) discuss the types of commonsense knowledge captured by modern NLP systems (e.g., large pretrained language models), and (4) present ways to measure systems' commonsense reasoning abilities. We will finish with (5) a discussion of various ways in which commonsense reasoning can be used to improve performance on NLP tasks, exemplified by an (6) interactive session on integrating commonsense into a downstream task.

Maarten Sap is a PhD student in the Paul G. Allen School of Computer Science & Engineering at the University of Washington. His research focuses primarily on social applications of NLP, specifically on endowing machines with social intelligence, social commonsense, or theory of mind.

Vered Shwartz is a postdoctoral researcher at the Allen Institute for Artificial Intelligence (AI2) and the Paul G. Allen School of Computer Science & Engineering at the University of Washington, working on lexical semantics, multiword expressions, and commonsense reasoning. She coorganized the ACL 2018 Student Research Workshop, the SemEval 2018 shared task on hypernymy discovery, and the AAAI 2020 Workshop on Reasoning for Complex Question Answering, Special Edition on Commonsense Reasoning.

Antoine Bosselut is a PhD student in the Paul G. Allen School of Computer Science & Engineering at the University of Washington and a student researcher at the Allen Institute for Artificial Intelligence (AI2). His research interests are in integrating commonsense knowledge and reasoning into downstream applications for text generation, summarization, and conversational dialogue. He organized the West Coast NLP (WeCNLP) in 2018 and 2019 and the NeuralGen workshop at NAACL 2019.

Yejin Choi is an associate professor at the Paul G. Allen School of Computer Science & Engineering at the University of Washington and also a senior research manager at AI2 overseeing the project Mosaic. Her research interests include language grounding with vision, physical and social commonsense knowledge, language generation with long-term coherence, conversational AI, and AI for social good. She was a recipient of Borg Early Career Award (BECA) in 2018, among the IEEEs AI Top 10 to Watch in 2015, a corecipient of the Marr Prize at ICCV 2013, and a faculty advisor for the Sounding Board team that won the inaugural Alexa Prize Challenge in 2017. She was on the steering committee of the NeuralGen workshop at NAACL 2019.

Dan Roth is the Eduardo D. Glandt Distinguished Professor at the Department of Computer and Information Science, University of Pennsylvania, and a Fellow of the AAAS, the ACM, AAAI, and the ACL. In 2017 Roth was awarded the John McCarthy Award, the highest award the AI community gives to mid-career AI researchers. He was the Editor-in-Chief of the Journal of Artificial Intelligence Research (JAIR) and a program co-chair of AAAI, ACL and CoNLL. Dan has presented several tutorials in conferences including at ACL, on entity linking, temporal reasoning, transferable representation learning, and more.

Tutorial 7

Integrating Ethics into the NLP Curriculum

Introductory

Emily M. Bender, Dirk Hovy, and Alexandra Schofield

[Website]

To raise awareness among future NLP practitioners and prevent inertia in the field, we need to place ethics in the curriculum for all NLP students—not as an elective, but as a core part of their education. Our goal in this tutorial is to empower NLP researchers and practitioners with tools and resources to teach others about how to ethically apply NLP techniques. We will present both high-level strategies for developing an ethics-oriented curriculum, based on experience and best practices, as well as specific sample exercises that can be brought to a classroom. This highly interactive work session will culminate in a shared online resource page that pools lesson plans, assignments, exercise ideas, reading suggestions, and ideas from the attendees. Though the tutorial will focus particularly on examples for university classrooms, we believe these ideas can extend to company-internal workshops or tutorials in a variety of organizations. In this setting, a key lesson is that there is no single approach to ethical NLP: each project requires thoughtful consideration about what steps can be taken to best support people affected by that project. However, we can learn (and teach) what issues to be aware of, what questions to ask, and what strategies are available to mitigate harm.

Emily M. Bender, University of Washington
ebender@uw.edu
faculty.washington.edu/ebender
Emily M. Bender is a Professor of Linguistics and Adjunct Professor of Computer Science and Engineering at the University of Washington. Her research interests include computational semantics, grammar engineering, computational linguistic typology, and ethics in NLP. She is the Faculty Director of UW's Professional Masters in Computational Linguistics (CLMS) and has been engaged with integrating ethics into the CLMS curriculum since 2016. She co-organized the first EthNLP workshop. Her first publication in this area is the TACL paper "Data Statements for NLP: Toward Mitigating System Bias and Enabling Better Science" (Bender and Friedman, 2018) and she has been an invited speaker at workshops and panels related to ethics and NLP (or AI more broadly) at the Taskar Memorial Event (UW, March 2018), The Future of Artificial Intelligence: Language, Ethics, Technology (Cambridge, March 2019), West Coast NLP (Facebook, September 2019), Machine Learning Competitions for All (NeurIPS, December 2019) and AAAS (Seattle, February 2020).

Xanda Schofield, Harvey Mudd College
xanda@cs.hmc.edu
www.cs.hmc.edu/ xanda
Xanda Schofield is an Assistant Professor of Computer Science at Harvey Mudd College. Her work focuses on the practical aspects of using distributional semantic models for analysis of realworld datasets, with problems ranging from understanding the consequences of data pre-processing on model inference (Schofield and Mimno, 2016; Schofield et al., 2017) to enforcing text privacy for these models (Schein et al., 2018). She also is interested in pedagogy at this intersection, having co-developed a Text Mining for History and Literature course at Cornell University with David Mimno. She is currently focusing pedagogical efforts on how to introduce considerations of ethics and bias into other courses such as Algorithms.

Dirk Hovy, Bocconi University
dirk.hovy@unibocconi.it
www.dirkhovy.com
Dirk Hovy is an Associate Professor of Computer Science in the Department of Marketing at Bocconi University in Milan, Italy. His research focuses on how social dimensions influence language and in turn NLP models, as well as on questions of bias and fairness. He strives to integrate sociolinguistic knowledge into NLP models to counteract demographic bias. Dirk has written on ethics and bias in NLP (Hovy and Spruit, 2016), co-organized two editions of the EthNLP workshops and one of the abusive language workshop, and was an invited speaker on panels on ethics at NAACL 2018 and SLT 2018. He is teaching a related tutorial (on ethic

Tutorial 8

Open-Domain Question Answering

Cutting-edge

Danqi Chen and Wen-tau Yih

[Website]

This tutorial provides a comprehensive and coherent overview of cutting-edge research in open-domain question answering (QA), the task of answering questions using a large collection of documents of diversified topics. We will start by first giving a brief historical background, discussing the basic setup and core technical challenges of the research problem, and then describe modern datasets with the common evaluation metrics and benchmarks. The focus will then shift to cutting-edge models proposed for open-domain QA, including two-stage retriever-reader approaches, dense retriever and end-to-end training, and retriever-free methods. Finally, we will cover some hybrid approaches using both text and large knowledge bases and conclude the tutorial with important open questions. We hope that the tutorial will not only help the audience to acquire up-to-date knowledge but also provide new perspectives to stimulate the advances of open-domain QA research in the next phase.

Danqi Chen Danqi Chen is an Assistant Professor of Computer Science at Princeton University and co-directs the Princeton NLP Group. Danqi's research interests lie within deep learning for natural language processing, with an emphasis on the intersection between text understanding and knowledge representation/reasoning and applications such as question answering and information extraction. Before joining Princeton University, Danqi worked as a visiting scientist at Facebook AI Research (FAIR). She received her PhD from Stanford University (advised by Christopher Manning) in 2018 and B.Eng from Tsinghua University in 2012.

Website: `https://www.cs.princeton.edu/ danqic/`

Scott Wen-tau Yih Scott Wen-tau Yih is a Research Scientist at Facebook AI Research (FAIR), and his recent research focuses on continuous representations and neural network models, with applications in knowledge base embedding, semantic parsing and question answering. Yih received the best paper award from CoNLL'11, an outstanding paper award from ACL'15 and has served as an area co-chair and a program co-chair for several top conferences. He is also a co-presenter for several popular tutorials on topics including Semantic Role Labeling, Deep Learning for NLP, Question Answering with Knowledge Base, Web and Beyond and NLP for Precision Medicine.

Website: `http://scottyih.org/`

3

Main Conference: Monday, July 6

Overview

5:00–5:45 **Demo Session 1A**

5:00–6:00 **Session 1A**
Cognitive Modeling and Psycholinguistics-1
Dialogue and Interactive Systems-1
Discourse and Pragmatics-1
Generation-1
Information Retrieval and Text Mining-1
Machine Translation-1
Student Research Workshop
Theory and Formalism in NLP (Linguistic and Mathematical)-1

5:45–6:30 **Demo Session 1B**

6:00–7:00 **Session 1B**
Computational Social Science and Social Media-1
Dialogue and Interactive Systems-2
Generation-2
Information Retrieval and Text Mining-2
NLP Applications-1
Question Answering-1
Resources and Evaluation-1
Lexical-1
Student Research Workshop

6:30–7:15 **Demo Session 1C**

8:00–8:45 **Demo Session 2A**

8:00–9:00 **Session 2A**
Computational Social Science and Social Media-2
Dialogue and Interactive Systems-3
Generation-3
Information Retrieval and Text Mining-3
Phonology, Morphology and Word Segmentation-1
Question Answering-2
Resources and Evaluation-2
Sentence Level-1
Student Research Workshop
Summarization-1

8:45–9:30 **Demo Session 2B**

9:00–10:00 **Session 2B**
Cognitive Modeling and Psycholinguistics-2
Dialogue and Interactive Systems-4
Discourse and Pragmatics-2
Generation-4
Information Extraction-1
Machine Translation-2
NLP Applications-2
Resources and Evaluation-3
Student Research Workshop
Theory and Formalism in NLP (Linguistic and Mathematical)-2

9:30–10:15 **Demo Session 2C**
12:00–12:45 **Demo Session 3A**
12:00–13:00 **Session 3A**
Cognitive Modeling and Psycholinguistics-3
Computational Social Science and Social Media-3
Dialogue and Interactive Systems-5
Generation-5
Information Retrieval and Text Mining-4
Machine Translation-3
Resources and Evaluation-4
Sentence Level-2
Student Research Workshop

12:45–13:30 **Demo Session 3B**
13:00–14:00 **Session 3B**
Dialogue and Interactive Systems-6
Discourse and Pragmatics-3
Generation-6
Information Extraction-2
Machine Translation-4
Phonology, Morphology and Word Segmentation-2
Student Research Workshop
Summarization-2

13:30–14:15 **Demo Session 3C**
14:00–14:15 **Opening Remarks**
14:15–14:30 **Presidential Address (Sponsored by Amazon Science and Baidu)**
14:30–15:15 **Keynote 1 Video Livestream: Kathleen R. McKeown (Sponsored by Facebook and Megagon Labs)**
15:15–15:45 **Keynote 1 Live Q&A: Kathleen R. McKeown (Sponsored by Fa cebook and Megagon Labs)**
15:45–16:15 **Business Meeting Q&A**
17:00–17:45 **Demo Session 4A**
17:00–18:00 **Session 4A**
Cognitive Modeling and Psycholinguistics-4
Dialogue and Interactive Systems-7
Machine Learning for NLP-1
NLP Applications-3
Question Answering-3
Textual Inference and Other Areas of Semantics-1
Speech and Multimodality-1
Student Research Workshop

17:45–18:30 **Demo Session 4B**
18:00–19:00 **Session 4B**
Dialogue and Interactive Systems-8
Generation-7
Language Grounding to Vision, Robotics and Beyond-1
Machine Learning for NLP-2
Machine Translation-5
Lexical-2
Student Research Workshop
Summarization-3

18:30–19:15 **Demo Session 4C**
20:00–20:45 **Demo Session 5A**

20:00–21:00 **Session 5A**
Dialogue and Interactive Systems-9
Generation-8
Information Retrieval and Text Mining-5
Machine Learning for NLP-3
Machine Translation-6
Textual Inference and Other Areas of Semantics-2
Speech and Multimodality-2
Theory and Formalism in NLP (Linguistic and Mathematical)-3

20:45–21:30 **Demo Session 5B**

21:00–22:00 **Session 5B**
Cognitive Modeling and Psycholinguistics-5
Dialogue and Interactive Systems-10
Language Grounding to Vision, Robotics and Beyond-2
Machine Learning for NLP-4
NLP Applications-4
Lexical-3

21:30–22:15 **Demo Session 5C**

Demo Session 1A

Time: *5:00–5:45*

Xiaomingbot: A Multilingual Robot News Reporter [Website][PDF]
Runxin Xu, Jun Cao, Mingxuan Wang, Jiaze Chen, Hao Zhou, Ying Zeng, Yuping Wang, Li Chen, Xiang Yin, Xijin Zhang, Songcheng Jiang, Yuxuan Wang, and Lei Li

This paper proposes the building of Xiaomingbot, an intelligent, multilingual and multimodal software robot equipped with four inte- gral capabilities: news generation, news translation, news reading and avatar animation. Its system summarizes Chinese news that it automatically generates from data tables. Next, it translates the summary or the full article into multiple languages, and reads the multi- lingual rendition through synthesized speech. Notably, Xiaomingbot utilizes a voice cloning technology to synthesize the speech trained from a real person's voice data in one input language. The proposed system enjoys several merits: it has an animated avatar, and is able to generate and read multilingual news. Since it was put into practice, Xiaomingbot has written over 600,000 articles, and gained over 150,000 followers on social media platforms.

Session 1A Overview – Monday, July 6, 2020 5:00–6:00

Track A *Cognitive Modeling and Psycholinguistics-1* Abstracts	[TACL] How Furiously Can Colourless Green Ideas Sleep? Sentence Acceptability in Context *Lau, Armendariz, Purver, Shu, and Lappin* [Website][PDF]	Learning to Understand Child-directed and Adult-directed Speech *Gelderloos, Chrupała, and Alishahi* [Website][PDF]	Predicting Depression in Screening Interviews from Latent Categorization of Interview Prompts *Rinaldi, Fox Tree, and Chaturvedi* [Website][PDF]		
Track B *Dialogue and Interactive Systems-1* Abstracts	Coach: A Coarse-to-Fine Approach for Cross-domain Slot Filling *Liu, Winata, Xu, and Fung* [Website][PDF]	Designing Precise and Robust Dialogue Response Evaluators *Zhao, Lala, and Kawahara* [Website][PDF]	Dialogue State Tracking with Explicit Slot Connection Modeling *Ouyang, Chen, Dai, Zhao, Huang, and CHEN* [Website][PDF]	Generating Informative Conversational Response using Recurrent Knowledge-Interaction and Knowledge-Copy *Lin, Jian, He, Wang, and Chu* [Website][PDF]	Guiding Variational Response Generator to Exploit Persona *Wu, Li, Wang, Chen, Wong, Huang, and Wang* [Website][PDF]
	Large Scale Multi-Actor Generative Dialog Modeling *Boyd, Puri, Shoeybi, Patwary, and Catanzaro* [Website][PDF]	PLATO: Pre-trained Dialogue Generation Model with Discrete Latent Variable *Bao, He, Wang, Wu, and Wang* [Website][PDF]	Slot-consistent NLG for Task-oriented Dialogue Systems with Iterative Rectification Network *Li, Yao, Qin, Che, Li, and Liu* [Website][PDF]	Span-ConveRT: Few-shot Span Extraction for Dialog with Pretrained Conversational Representations *Coope, Farghly, Gerz, Vulić, and Henderson* [Website][PDF]	Zero-Shot Transfer Learning with Synthesized Data for Multi-Domain Dialogue State Tracking *Campagna, Foryciarz, Moradshahi, and Lam* [Website][PDF]
Track C *Discourse and Pragmatics-1* Abstracts	A Complete Shift-Reduce Chinese Discourse Parser with Robust Dynamic Oracle *Hung, Huang, and Chen* [Website][PDF]	TransS-Driven Joint Learning Architecture for Implicit Discourse Relation Recognition *He, Wang, Guo, and Han* [Website][PDF]			
Track D *Generation-1* Abstracts	A Study of Non-autoregressive Model for Sequence Generation *Ren, Liu, Tan, Zhao, and Liu* [Website][PDF]	Cross-modal Language Generation using Pivot Stabilization for Web-scale Language Coverage *Thapliyal and Soricut* [Website][PDF]	Fact-based Text Editing *Iso, Qiao, and Li* [Website][PDF]	Few-Shot NLG with Pre-Trained Language Model *Chen, Eavani, Chen, Liu, and Wang* [Website][PDF]	Fluent Response Generation for Conversational Question Answering *Baheti, Ritter, and Small* [Website][PDF]
	Generating Diverse and Consistent QA pairs from Contexts with Information-Maximizing Hierarchical Conditional VAEs *Lee, Lee, Jeong, Kim, and Hwang* [Website][PDF]	Learning to Ask More: Semi-Autoregressive Sequential Question Generation under Dual-Graph Interaction *Chai and Wan* [Website][PDF]	Neural Syntactic Preordering for Controlled Paraphrase Generation *Goyal and Durrett* [Website][PDF]	Pre-train and Plug-in: Flexible Conditional Text Generation with Variational Auto-Encoders *Duan, Xu, Pei, Han, and Li* [Website][PDF]	Probabilistically Masked Language Model Capable of Autoregressive Generation in Arbitrary Word Order *Liao, Jiang, and Liu* [Website][PDF]

	Reverse Engineering Configurations of Neural Text Generation Models *Tay, Bahri, Zheng, Brunk, Metzler, and Tomkins* [Website][PDF]	Review-based Question Generation with Adaptive Instance Transfer and Augmentation *Yu, Bing, Zhang, Lam, and Si* [Website][PDF]	TAG : Type Auxiliary Guiding for Code Comment Generation *Cai, Liang, Xu, Hao, and Chen* [Website][PDF]	Unsupervised Paraphrasing by Simulated Annealing *Liu, Mou, Meng, Zhou, Zhou, and Song* [Website][PDF]	
Track E *Information Retrieval and Text Mining-1* Abstracts	A Joint Model for Document Segmentation and Segment Labeling *Barrow, Jain, Morariu, Manjunatha, Oard, and Resnik* [Website][PDF]	Contextualized Weak Supervision for Text Classification *Mekala and Shang* [Website][PDF]	Every Document Owns Its Structure: Inductive Text Classification via Graph Neural Networks *Zhang, Yu, Cui, Wu, Wen, and Wang* [Website][PDF]	Neural Topic Modeling with Bidirectional Adversarial Training *Wang, Hu, Zhou, He, Xiong, Ye, and Xu* [Website][PDF]	Text Classification with Negative Supervision *Ohashi, Takayama, Kajiwara, Chu, and Arase* [Website][PDF]
Track F *Machine Translation-1* Abstracts	Content Word Aware Neural Machine Translation *Chen, Wang, Utiyama, and Sumita* [Website][PDF]	Evaluating Explanation Methods for Neural Machine Translation *Li, Liu, Li, Li, Huang, and Shi* [Website][PDF]	Jointly Masked Sequence-to-Sequence Model for Non-Autoregressive Neural Machine Translation *Guo, Xu, and Chen* [Website][PDF]	Learning Source Phrase Representations for Neural Machine Translation *Xu, Genabith, Xiong, Liu, and Zhang* [Website][PDF]	Lipschitz Constrained Parameter Initialization for Deep Transformers *Xu, Liu, Genabith, Xiong, and Zhang* [Website][PDF]
	Location Attention for Extrapolation to Longer Sequences *Dubois, Dagan, Hupkes, and Bruni* [Website][PDF]	Multiscale Collaborative Deep Models for Neural Machine Translation *Wei, Yu, Hu, Zhang, Weng, and Luo* [Website][PDF]	Norm-Based Curriculum Learning for Neural Machine Translation *Liu, Lai, Wong, and Chao* [Website][PDF]	Opportunistic Decoding with Timely Correction for Simultaneous Translation *Zheng, Ma, Zheng, Liu, and Huang* [Website][PDF]	
Track G *Student Research Workshop* Abstracts	Adaptive Transformers for Learning Multimodal Representations *Bhargava* [Website][PDF]	Story-level Text Style Transfer: A Proposal *Qian* [Website][PDF]	Unsupervised Paraphasia Classification in Aphasic Speech *Pai, Sachdeva, Sachdeva, and Shah* [Website][PDF]	HGCN4MeSH: Hybrid Graph Convolution Network for MeSH Indexing *Yu, Yang, and Li* [Website][PDF]	
Track H *Theory and Formalism in NLP (Linguistic and Mathematical)-1* Abstracts	A Formal Hierarchy of RNN Architectures *Merrill, Weiss, Goldberg, Schwartz, Smith, and Yahav* [Website][PDF]	A Three-Parameter Rank-Frequency Relation in Natural Languages *Ding, Utiyama, and Sumita* [Website][PDF]	Dice Loss for Data-imbalanced NLP Tasks *Li, Sun, Meng, Liang, Wu, and Li* [Website][PDF]	Emergence of Syntax Needs Minimal Supervision *Bailly and Gábor* [Website][PDF]	Language Models as an Alternative Evaluator of Word Order Hypotheses: A Case Study in Japanese *Kuribayashi, Ito, Suzuki, and Inui* [Website][PDF]
	[TACL] Theoretical Limitations of Self-Attention in Neural Sequence Models *Hahn* [Website][PDF]				

Session 1A Details

Session 1A: Cognitive Modeling and Psycholinguistics-1

[TACL] How Furiously Can Colourless Green Ideas Sleep? Sentence Acceptability in Context [Website][PDF]
Jey Han Lau, Carlos Santos Armendariz, Matthew Purver, Chang Shu, and Shalom Lappin 5:00–6:00

We study the influence of context on sentence acceptability. First we compare the acceptability ratings of sentences judged in isolation, with a relevant context, and with an irrelevant context. Our results show that context induces a cognitive load for humans, which compresses the distribution of ratings. Moreover, in relevant contexts we observe a discourse coherence effect which uniformly raises acceptability. Next, we test unidirectional and bidirectional language models in their ability to predict acceptability ratings. The bidirectional models show very promising results, with the best model achieving a new state-of-the-art for unsupervised acceptability prediction. The two sets of experiments provide insights into the cognitive aspects of sentence processing and central issues in the computational modelling of text and discourse.

Learning to Understand Child-directed and Adult-directed Speech [Website][PDF]
Lieke Gelderloos, Grzegorz Chrupała, and Afra Alishahi 5:00–6:00

Speech directed to children differs from adult-directed speech in linguistic aspects such as repetition, word choice, and sentence length, as well as in aspects of the speech signal itself, such as prosodic and phonemic variation. Human language acquisition research indicates that child-directed speech helps language learners. This study explores the effect of child-directed speech when learning to extract semantic information from speech directly. We compare the task performance of models trained on adult-directed speech (ADS) and child-directed speech (CDS). We find indications that CDS helps in the initial stages of learning, but eventually, models trained on ADS reach comparable task performance, and generalize better. The results suggest that this is at least partially due to linguistic rather than acoustic properties of the two registers, as we see the same pattern when looking at models trained on acoustically comparable synthetic speech.

Predicting Depression in Screening Interviews from Latent Categorization of Interview Prompts [Website][PDF]
Alex Rinaldi, Jean Fox Tree, and Snigdha Chaturvedi 5:00–6:00

Accurately diagnosing depression is difficult– requiring time-intensive interviews, assessments, and analysis. Hence, automated methods that can assess linguistic patterns in these interviews could help psychiatric professionals make faster, more informed decisions about diagnosis. We propose JLPC, a model that analyzes interview transcripts to identify depression while jointly categorizing interview prompts into latent categories. This latent categorization allows the model to define high-level conversational contexts that influence patterns of language in depressed individuals. We show that the proposed model not only outperforms competitive baselines, but that its latent prompt categories provide psycholinguistic insights about depression.

Session 1A: Dialogue and Interactive Systems-1

Coach: A Coarse-to-Fine Approach for Cross-domain Slot Filling [Website][PDF]
Zihan Liu, Genta Indra Winata, Peng Xu, and Pascale Fung 5:00–6:00

As an essential task in task-oriented dialog systems, slot filling requires extensive training data in a certain domain. However, such data are not always available. Hence, cross-domain slot filling has naturally arisen to cope with this data scarcity problem. In this paper, we propose a Coarse-to-fine approach (Coach) for cross-domain slot filling. Our model first learns the general pattern of slot entities by detecting whether the tokens are slot entities or not. It then predicts the specific types for the slot entities. In addition, we propose a template regularization approach to improve the adaptation robustness by regularizing the representation of utterances based on utterance templates. Experimental results show that our model significantly outperforms state-of-the-art approaches in slot filling. Furthermore, our model can also be applied to the cross-domain named entity recognition task, and it achieves better adaptation performance than other existing baselines. The code is available at https://github.com/zliucr/coach.

Designing Precise and Robust Dialogue Response Evaluators [Website][PDF]
Tianyu Zhao, Divesh Lala, and Tatsuya Kawahara 5:00–6:00

Automatic dialogue response evaluator has been proposed as an alternative to automated metrics and human evaluation. However, existing automatic evaluators achieve only moderate correlation with human judgement and they are not robust. In this work, we propose to build a reference-free evaluator and exploit the power of semi-supervised training and pretrained (masked) language models. Experimental results demonstrate that the proposed evaluator achieves a strong correlation (> 0.6) with human judgement and generalizes robustly to diverse responses and corpora. We open-source the code and data in https://github.com/ZHAOTING/dialog-processing.

Dialogue State Tracking with Explicit Slot Connection Modeling [Website][PDF]
Yawen Ouyang, Moxin Chen, Xinyu Dai, Yinggong Zhao, Shujian Huang, and Jiajun CHEN 5:00–6:00

Recent proposed approaches have made promising progress in dialogue state tracking (DST). However, in multi-domain scenarios, ellipsis and reference are frequently adopted by users to express values that have been mentioned by slots from other domains. To handle these phenomena, we propose a Dialogue State Tracking with Slot Connections (DST-SC) model to explicitly consider slot correlations across different domains. Given a target slot, the slot connecting mechanism in DST-SC can infer its source slot and copy the source slot value directly, thus significantly reducing the difficulty of learning and reasoning. Experimental results verify the benefits of explicit slot connection modeling, and our model achieves state-of-the-art performance on MultiWOZ 2.0 and MultiWOZ 2.1 datasets.

Generating Informative Conversational Response using Recurrent Knowledge-Interaction and Knowledge-Copy [Website][PDF]
Xiexiong Lin, Weiyu Jian, Jianshan He, Taifeng Wang, and Wei Chu 5:00–6:00

Knowledge-driven conversation approaches have achieved remarkable research attention recently. However, generating an informative response with multiple relevant knowledge without losing fluency and coherence is still one of the main challenges. To address this issue, this paper proposes a method that uses recurrent knowledge interaction among response decoding steps to incorporate appropriate knowledge. Furthermore, we introduce a knowledge copy mechanism using a knowledge-aware pointer network to copy words from external knowledge according to knowledge attention distribution. Our joint neural conversation model which integrates recurrent Knowledge-Interaction and knowledge Copy (KIC) performs well on generating informative responses. Experiments demonstrate that our model with fewer parameters yields significant improvements over competitive baselines on two datasets Wizard-of-Wikipedia(average Bleu +87%; abs.: 0.034) and DuConv(average Bleu +20%; abs.: 0.047)) with different knowledge formats (textual & structured) and different languages (English & Chinese).

Guiding Variational Response Generator to Exploit Persona [Website][PDF]
Bowen Wu, MengYuan Li, Zongsheng Wang, Yifu Chen, Derek F. Wong, qihang feng qihang, Junhong Huang, and Baoxun Wang 5:00–6:00

Leveraging persona information of users in Neural Response Generators (NRG) to perform personalized conversations has been considered as an attractive and important topic in the research of conversational agents over the past few years. Despite of the promising progress achieved by recent studies in this field, persona information tends to be incorporated into neural networks in the form of user embeddings, with the expectation that the persona can be involved via End-to-End learning. This paper proposes to adopt the personality-related characteristics of human conversations into variational response generators, by designing a specific conditional variational autoencoder based deep model with two new regularization terms employed to the loss function, so as to guide the optimization towards the direction of generating both persona-aware and relevant responses. Besides, to reasonably evaluate the performances of various persona modeling approaches, this paper further presents three direct persona-oriented metrics from different perspectives. The experimental results have shown that our proposed methodology can notably improve the performance of persona-aware response generation, and the metrics are reasonable to evaluate the results.

Large Scale Multi-Actor Generative Dialog Modeling [Website][PDF]
Alex Boyd, Raul Puri, Mohammad Shoeybi, Mostofa Patwary, and Bryan Catanzaro 5:00–6:00

Non-goal oriented dialog agents (i.e. chatbots) aim to produce varying and engaging conversations with a user; however, they typically exhibit either inconsistent personality across conversations or the average personality of all users. This paper addresses these issues by controlling an agent's persona upon generation via conditioning on prior conversations of a target actor. In doing so, we are able to utilize more abstract patterns within a person's speech and

better emulate them in generated responses. This work introduces the Generative Conversation Control model, an augmented and fine-tuned GPT-2 language model that conditions on past reference conversations to probabilistically model multi-turn conversations in the actor's persona. We introduce an accompanying data collection procedure to obtain 10.3M conversations from 6 months worth of Reddit comments. We demonstrate that scaling model sizes from 117M to 8.3B parameters yields an improvement from 23.14 to 13.14 perplexity on 1.7M held out Reddit conversations. Increasing model scale yielded similar improvements in human evaluations that measure preference of model samples to the held out target distribution in terms of realism (31% increased to 37% preference), style matching (37% to 42%), grammar and content quality (29% to 42%), and conversation coherency (32% to 40%). We find that conditionally modeling past conversations improves perplexity by 0.47 in automatic evaluations. Through human trials we identify positive trends between conditional modeling and style matching and outline steps to further improve persona control.

PLATO: Pre-trained Dialogue Generation Model with Discrete Latent Variable [Website][PDF]
Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang 5:00–6:00

Pre-training models have been proved effective for a wide range of natural language processing tasks. Inspired by this, we propose a novel dialogue generation pre-training framework to support various kinds of conversations, including chit-chat, knowledge grounded dialogues, and conversational question answering. In this framework, we adopt flexible attention mechanisms to fully leverage the bi-directional context and the uni-directional characteristic of language generation. We also introduce discrete latent variables to tackle the inherent one-to-many mapping problem in response generation. Two reciprocal tasks of response generation and latent act recognition are designed and carried out simultaneously within a shared network. Comprehensive experiments on three publicly available datasets verify the effectiveness and superiority of the proposed framework.

Slot-consistent NLG for Task-oriented Dialogue Systems with Iterative Rectification Network [Website][PDF]
Yangming Li, Kaisheng Yao, Libo Qin, Wanxiang Che, Xiaolong Li, and Ting Liu 5:00–6:00

Data-driven approaches using neural networks have achieved promising performances in natural language generation (NLG). However, neural generators are prone to make mistakes, e.g., neglecting an input slot value and generating a redundant slot value. Prior works refer this to hallucination phenomenon. In this paper, we study slot consistency for building reliable NLG systems with all slot values of input dialogue act (DA) properly generated in output sentences. We propose Iterative Rectification Network (IRN) for improving general NLG systems to produce both correct and fluent responses. It applies a bootstrapping algorithm to sample training candidates and uses reinforcement learning to incorporate discrete reward related to slot inconsistency into training. Comprehensive studies have been conducted on multiple benchmark datasets, showing that the proposed methods have significantly reduced the slot error rate (ERR) for all strong baselines. Human evaluations also have confirmed its effectiveness.

Span-ConveRT: Few-shot Span Extraction for Dialog with Pretrained Conversational Representations [Website][PDF]
Samuel Coope, Tyler Farghly, Daniela Gerz, Ivan Vulić, and Matthew Henderson 5:00–6:00

We introduce Span-ConveRT, a light-weight model for dialog slot-filling which frames the task as a turn-based span extraction task. This formulation allows for a simple integration of conversational knowledge coded in large pretrained conversational models such as ConveRT (Henderson et al., 2019). We show that leveraging such knowledge in Span-ConveRT is especially useful for few-shot learning scenarios: we report consistent gains over 1) a span extractor that trains representations from scratch in the target domain, and 2) a BERT-based span extractor. In order to inspire more work on span extraction for the slot-filling task, we also release RESTAURANTS-8K, a new challenging data set of 8,198 utterances, compiled from actual conversations in the restaurant booking domain.

Zero-Shot Transfer Learning with Synthesized Data for Multi-Domain Dialogue State Tracking [Website][PDF]
Giovanni Campagna, Agata Foryciarz, Mehrad Moradshahi, and Monica Lam 5:00–6:00

Zero-shot transfer learning for multi-domain dialogue state tracking can allow us to handle new domains without incurring the high cost of data acquisition. This paper proposes new zero-short transfer learning technique for dialogue state tracking where the in-domain training data are all synthesized from an abstract dialogue model and the ontology of the domain. We show that data augmentation through synthesized data can improve the accuracy of zero-shot learning for both the TRADE model and the BERT-based SUMBT model on the MultiWOZ 2.1 dataset. We show training with only synthesized in-domain data on the SUMBT model can reach about 2/3 of the accuracy obtained with the full training dataset. We improve the zero-shot learning state of the art on average across domains by 21%.

Session 1A: Discourse and Pragmatics-1

A Complete Shift-Reduce Chinese Discourse Parser with Robust Dynamic Oracle [Website][PDF]
Shyh-Shiun Hung, Hen-Hsen Huang, and Hsin-Hsi Chen 5:00–6:00

This work proposes a standalone, complete Chinese discourse parser for practical applications. We approach Chinese discourse parsing from a variety of aspects and improve the shift-reduce parser not only by integrating the pre-trained text encoder, but also by employing novel training strategies. We revise the dynamic-oracle procedure for training the shift-reduce parser, and apply unsupervised data augmentation to enhance rhetorical relation recognition. Experimental results show that our Chinese discourse parser achieves the state-of-the-art performance.

TransS-Driven Joint Learning Architecture for Implicit Discourse Relation Recognition [Website][PDF]
Ruifang He, Jian Wang, Fengyu Guo, and Yugui Han 5:00–6:00

Implicit discourse relation recognition is a challenging task due to the lack of connectives as strong linguistic clues. Previous methods primarily encode two arguments separately or extract the specific interaction patterns for the task, which have not fully exploited the annotated relation signal. Therefore, we propose a novel TransS-driven joint learning architecture to address the issues. Specifically, based on the multi-level encoder, we 1) translate discourse relations in low-dimensional embedding space (called TransS), which could mine the latent geometric structure information of argument-relation instances; 2) further exploit the semantic features of arguments to assist discourse understanding; 3) jointly learn 1) and 2) to mutually reinforce each other to obtain the better argument representations, so as to improve the performance of the task. Extensive experimental results on the Penn Discourse TreeBank (PDTB) show that our model achieves competitive results against several state-of-the-art systems.

Session 1A: Generation-1

A Study of Non-autoregressive Model for Sequence Generation [Website][PDF]
Yi Ren, Jinglin Liu, Xu Tan, Zhou Zhao, sheng zhao sheng, and Tie-Yan Liu 5:00–6:00

Non-autoregressive (NAR) models generate all the tokens of a sequence in parallel, resulting in faster generation speed compared to their autoregressive (AR) counterparts but at the cost of lower accuracy. Different techniques including knowledge distillation and source-target alignment have been proposed to bridge the gap between AR and NAR models in various tasks such as neural machine translation (NMT), automatic speech recognition (ASR), and text to speech (TTS). With the help of those techniques, NAR models can catch up with the accuracy of AR models in some tasks but not in some others. In this work, we conduct a study to understand the difficulty of NAR sequence generation and try to answer: (1) Why NAR models can catch up with AR models in some tasks but not all? (2) Why techniques like knowledge distillation and source-target alignment can help NAR models. Since the main difference between AR and NAR models is that NAR models do not use dependency among target tokens while AR models do, intuitively the difficulty of NAR sequence generation heavily depends on the strongness of dependency among target tokens. To quantify such dependency, we propose an analysis model called CoMMA to characterize the difficulty of different NAR sequence generation tasks. We have several interesting findings: 1) Among the NMT, ASR and TTS tasks, ASR has the most target-token dependency while TTS has the least. 2) Knowledge distillation reduces the target-token dependency in target sequence and thus improves the accuracy of NAR models. 3) Source-target alignment constraint encourages dependency of a target token on source tokens and thus eases the training of NAR models.

Cross-modal Language Generation using Pivot Stabilization for Web-scale Language Coverage [Website][PDF]
Ashish V. Thapliyal and Radu Soricut 5:00–6:00

Cross-modal language generation tasks such as image captioning are directly hurt in their ability to support non-English languages by the trend of data-hungry models combined with the lack of non-English annotations. We investigate potential solutions for combining existing language-generation annotations in English with translation capabilities in order to create solutions at web-scale in both domain and language coverage. We describe an approach called Pivot-Language Generation Stabilization (PLuGS), which leverages directly at training time both existing English annotations (gold data) as well as their machine-translated versions (silver data); at run-time, it generates first an English caption and then a corresponding target-language caption. We show that PLuGS models outperform other candidate solutions in evaluations performed over 5 different target languages, under a large-domain testset using images from the Open Images dataset. Furthermore, we find an interesting effect where the English captions generated by the PLuGS models are better than the captions generated by the original, monolingual English model.

Fact-based Text Editing [Website][PDF]
Hayate Iso, Chao Qiao, and Hang Li 5:00–6:00

We propose a novel text editing task, referred to as *fact-based text editing*, in which the goal is to revise a given document to better describe the facts in a knowledge base (e.g., several triples). The task is important in practice because reflecting the truth is a common requirement in text editing. First, we propose a method for automatically generating a dataset for research on fact-based text editing, where each instance consists of a draft text, a revised text, and several facts represented in triples. We apply the method into two public table-to-text datasets, obtaining two new datasets consisting of 233k and 37k instances, respectively. Next, we propose a new neural network architecture for fact-based text editing, called FACTEDITOR, which edits a draft text by referring to given facts using a buffer, a stream, and a memory. A straightforward approach to address the problem would be to employ an encoder-decoder model. Our experimental results on the two datasets show that FACTEDITOR outperforms the encoder-decoder approach in terms of fidelity and fluency. The results also show that FACTEDITOR conducts inference faster than the encoder-decoder approach.

Few-Shot NLG with Pre-Trained Language Model [Website][PDF]
Zhiyu Chen, Harini Eavani, Wenhu Chen, Yinyin Liu, and William Yang Wang 5:00–6:00

Neural-based end-to-end approaches to natural language generation (NLG) from structured data or knowledge are data-hungry, making their adoption for real-world applications difficult with limited data. In this work, we propose the new task of few-shot natural language generation. Motivated by how humans tend to summarize tabular data, we propose a simple yet effective approach and show that it not only demonstrates strong performance but also provides good generalization across domains. The design of the model architecture is based on two aspects: content selection from input data and language modeling to compose coherent sentences, which can be acquired from prior knowledge. With just 200 training examples, across multiple domains, we show that our approach achieves very reasonable performances and outperforms the strongest baseline by an average of over 8.0 BLEU points improvement. Our code and data can be found at https://github.com/czyssrs/Few-Shot-NLG

Fluent Response Generation for Conversational Question Answering [Website][PDF]
Ashutosh Baheti, Alan Ritter, and Kevin Small 5:00–6:00

Question answering (QA) is an important aspect of open-domain conversational agents, garnering specific research focus in the conversational QA (ConvQA) subtask. One notable limitation of recent ConvQA efforts is the response being answer span extraction from the target corpus, thus ignoring the natural language generation (NLG) aspect of high-quality conversational agents. In this work, we propose a method for situating QA responses within a SEQ2SEQ NLG approach to generate fluent grammatical answer responses while maintaining correctness. From a technical perspective, we use data augmentation to generate training data for an end-to-end system. Specifically, we develop

Syntactic Transformations (STs) to produce question-specific candidate answer responses and rank them using a BERT-based classifier (Devlin et al., 2019). Human evaluation on SQuAD 2.0 data (Rajpurkar et al., 2018) demonstrate that the proposed model outperforms baseline CoQA and QuAC models in generating conversational responses. We further show our model's scalability by conducting tests on the CoQA dataset. The code and data are available at https://github.com/abaheti95/QADialogSystem.

Generating Diverse and Consistent QA pairs from Contexts with Information-Maximizing Hierarchical Conditional VAEs [Website][PDF]
Dong Bok Lee, Seanie Lee, Woo Tae Jeong, Donghwan Kim, and Sung Ju Hwang 5:00–6:00

One of the most crucial challenges in question answering (QA) is the scarcity of labeled data, since it is costly to obtain question-answer (QA) pairs for a target text domain with human annotation. An alternative approach to tackle the problem is to use automatically generated QA pairs from either the problem context or from large amount of unstructured texts (e.g. Wikipedia). In this work, we propose a hierarchical conditional variational autoencoder (HCVAE) for generating QA pairs given unstructured texts as contexts, while maximizing the mutual information between generated QA pairs to ensure their consistency. We validate our Information Maximizing Hierarchical Conditional Variational AutoEncoder (Info-HCVAE) on several benchmark datasets by evaluating the performance of the QA model (BERT-base) using only the generated QA pairs (QA-based evaluation) or by using both the generated and human-labeled pairs (semi-supervised learning) for training, against state-of-the-art baseline models. The results show that our model obtains impressive performance gains over all baselines on both tasks, using only a fraction of data for training.

Learning to Ask More: Semi-Autoregressive Sequential Question Generation under Dual-Graph Interaction [Website][PDF]
Zi Chai and Xiaojun Wan 5:00–6:00

Traditional Question Generation (TQG) aims to generate a question given an input passage and an answer. When there is a sequence of answers, we can perform Sequential Question Generation (SQG) to produce a series of interconnected questions. Since the frequently occurred information omission and coreference between questions, SQG is rather challenging. Prior works regarded SQG as a dialog generation task and recurrently produced each question. However, they suffered from problems caused by error cascades and could only capture limited context dependencies. To this end, we generate questions in a semi-autoregressive way. Our model divides questions into different groups and generates each group of them in parallel. During this process, it builds two graphs focusing on information from passages, answers respectively and performs dual-graph interaction to get information for generation. Besides, we design an answer-aware attention mechanism and the coarse-to-fine generation scenario. Experiments on our new dataset containing 81.9K questions show that our model substantially outperforms prior works.

Neural Syntactic Preordering for Controlled Paraphrase Generation [Website][PDF]
Tanya Goyal and Greg Durrett 5:00–6:00

Paraphrasing natural language sentences is a multifaceted process: it might involve replacing individual words or short phrases, local rearrangement of content, or high-level restructuring like topicalization or passivization. Past approaches struggle to cover this space of paraphrase possibilities in an interpretable manner. Our work, inspired by pre-ordering literature in machine translation, uses syntactic transformations to softly "reorder" the source sentence and guide our neural paraphrasing model. First, given an input sentence, we derive a set of feasible syntactic rearrangements using an encoder-decoder model. This model operates over a partially lexical, partially syntactic view of the sentence and can reorder big chunks. Next, we use each proposed rearrangement to produce a sequence of position embeddings, which encourages our final encoder-decoder paraphrase model to attend to the source words in a particular order. Our evaluation, both automatic and human, shows that the proposed system retains the quality of the baseline approaches while giving a substantial increase in the diversity of the generated paraphrases.

Pre-train and Plug-in: Flexible Conditional Text Generation with Variational Auto-Encoders [Website][PDF]
Yu Duan, Canwen Xu, Jiaxin Pei, Jialong Han, and Chenliang Li 5:00–6:00

Conditional Text Generation has drawn much attention as a topic of Natural Language Generation (NLG) which provides the possibility for humans to control the properties of generated contents. Current conditional generation models cannot handle emerging conditions due to their joint end-to-end learning fashion. When a new condition added, these techniques require full retraining. In this paper, we present a new framework named Pre-train and Plug-in Variational Auto-Encoder (PPVAE) towards flexible conditional text generation. PPVAE decouples the text generation module from the condition representation module to allow "one-to-many" conditional generation. When a fresh condition emerges, only a lightweight network needs to be trained and works as a plug-in for PPVAE, which is efficient and desirable for real-world applications. Extensive experiments demonstrate the superiority of PPVAE against the existing alternatives with better conditionality and diversity but less training effort.

Probabilistically Masked Language Model Capable of Autoregressive Generation in Arbitrary Word Order [Website][PDF]
Yi Liao, Xin Jiang, and Qun Liu 5:00–6:00

Masked language model and autoregressive language model are two types of language models. While pretrained masked language models such as BERT overwhelm the line of natural language understanding (NLU) tasks, autoregressive language models such as GPT are especially capable in natural language generation (NLG). In this paper, we propose a probabilistic masking scheme for the masked language model, which we call probabilistically masked language model (PMLM). We implement a specific PMLM with a uniform prior distribution on the masking ratio named

u-PMLM. We prove that u-PMLM is equivalent to an autoregressive permutated language model. One main advantage of the model is that it supports text generation in arbitrary order with surprisingly good quality, which could potentially enable new applications over traditional unidirectional generation. Besides, the pretrained u-PMLM also outperforms BERT on a bunch of downstream NLU tasks.

Reverse Engineering Configurations of Neural Text Generation Models [Website][PDF]
Yi Tay, Dara Bahri, Che Zheng, Clifford Brunk, Donald Metzler, and Andrew Tomkins 5:00–6:00

Recent advances in neural text generation modeling have resulted in a number of societal concerns related to how such approaches might be used in malicious ways. It is therefore desirable to develop a deeper understanding of the fundamental properties of such models. The study of artifacts that emerge in machine generated text as a result of modeling choices is a nascent research area. To this end, the extent and degree to which these artifacts surface in generated text is still unclear. In the spirit of better understanding generative text models and their artifacts, we propose the new task of distinguishing which of several variants of a given model generated some piece of text. Specifically, we conduct an extensive suite of diagnostic tests to observe whether modeling choices (e.g., sampling methods, top-k probabilities, model architectures, etc.) leave detectable artifacts in the text they generate. Our key finding, which is backed by a rigorous set of experiments, is that such artifacts are present and that different modeling choices can be inferred by looking at generated text alone. This suggests that neural text generators may actually be more sensitive to various modeling choices than previously thought.

Review-based Question Generation with Adaptive Instance Transfer and Augmentation [Website][PDF]
Qian Yu, Lidong Bing, Qiong Zhang, Wai Lam, and Luo Si 5:00–6:00

While online reviews of products and services become an important information source, it remains inefficient for potential consumers to exploit verbose reviews for fulfilling their information need. We propose to explore question generation as a new way of review information exploitation, namely generating questions that can be answered by the corresponding review sentences. One major challenge of this generation task is the lack of training data, i.e. explicit mapping relation between the user-posed questions and review sentences. To obtain proper training instances for the generation model, we propose an iterative learning framework with adaptive instance transfer and augmentation. To generate to the point questions about the major aspects in reviews, related features extracted in an unsupervised manner are incorporated without the burden of aspect annotation. Experiments on data from various categories of a popular E-commerce site demonstrate the effectiveness of the framework, as well as the potentials of the proposed review-based question generation task.

TAG : Type Auxiliary Guiding for Code Comment Generation [Website][PDF]
Ruichu Cai, Zhihao Liang, Boyan Xu, zijian li zijian, Yuexing Hao, and Yao Chen 5:00–6:00

Existing leading code comment generation approaches with the structure-to-sequence framework ignores the type information of the interpretation of the code, e.g., operator, string, etc. However, introducing the type information into the existing framework is non-trivial due to the hierarchical dependence among the type information. In order to address the issues above, we propose a Type Auxiliary Guiding encoder-decoder framework for the code comment generation task which considers the source code as an N-ary tree with type information associated with each node. Specifically, our framework is featured with a Type-associated Encoder and a Type-restricted Decoder which enables adaptive summarization of the source code. We further propose a hierarchical reinforcement learning method to resolve the training difficulties of our proposed framework. Extensive evaluations demonstrate the state-of-the-art performance of our framework with both the auto-evaluated metrics and case studies.

Unsupervised Paraphrasing by Simulated Annealing [Website][PDF]
Xianggen Liu, Lili Mou, Fandong Meng, Hao Zhou, Jie Zhou, and Sen Song 5:00–6:00

We propose UPSA, a novel approach that accomplishes Unsupervised Paraphrasing by Simulated Annealing. We model paraphrase generation as an optimization problem and propose a sophisticated objective function, involving semantic similarity, expression diversity, and language fluency of paraphrases. UPSA searches the sentence space towards this objective by performing a sequence of local editing. We evaluate our approach on various datasets, namely, Quora, Wikianswers, MSCOCO, and Twitter. Extensive results show that UPSA achieves the state-of-the-art performance compared with previous unsupervised methods in terms of both automatic and human evaluations. Further, our approach outperforms most existing domain-adapted supervised models, showing the generalizability of UPSA.

Session 1A: Information Retrieval and Text Mining-1

A Joint Model for Document Segmentation and Segment Labeling [Website][PDF]
Joe Barrow, Rajiv Jain, Vlad Morariu, Varun Manjunatha, Douglas Oard, and Philip Resnik 5:00–6:00

Text segmentation aims to uncover latent structure by dividing text from a document into coherent sections. Where previous work on text segmentation considers the tasks of document segmentation and segment labeling separately, we show that the tasks contain complementary information and are best addressed jointly. We introduce Segment Pooling LSTM (S-LSTM), which is capable of jointly segmenting a document and labeling segments. In support of joint training, we develop a method for teaching the model to recover from errors by aligning the predicted and ground truth segments. We show that S-LSTM reduces segmentation error by 30% on average, while also improving segment labeling.

Contextualized Weak Supervision for Text Classification [Website][PDF]
Dheeraj Mekala and Jingbo Shang 5:00–6:00

Weakly supervised text classification based on a few user-provided seed words has recently attracted much attention from researchers. Existing methods mainly generate pseudo-labels in a context-free manner (e.g., string matching), therefore, the ambiguous, context-dependent nature of human language has been long overlooked. In this paper, we propose a novel framework ConWea, providing contextualized weak supervision for text classification. Specifically, we leverage contextualized representations of word occurrences and seed word information to automatically differentiate multiple interpretations of the same word, and thus create a contextualized corpus. This contextualized corpus is further utilized to train the classifier and expand seed words in an iterative manner. This process not only adds new contextualized, highly label-indicative keywords but also disambiguates initial seed words, making our weak supervision fully contextualized. Extensive experiments and case studies on real-world datasets demonstrate the necessity and significant advantages of using contextualized weak supervision, especially when the class labels are fine-grained.

Every Document Owns Its Structure: Inductive Text Classification via Graph Neural Networks [Website][PDF]
Yufeng Zhang, Xueli Yu, Zeyu Cui, Shu Wu, Zhongzhen Wen, and Liang Wang 5:00–6:00

Text classification is fundamental in natural language processing (NLP) and Graph Neural Networks (GNN) are recently applied in this task. However, the existing graph-based works can neither capture the contextual word relationships within each document nor fulfil the inductive learning of new words. Therefore in this work, to overcome such problems, we propose TextING for inductive text classification via GNN. We first build individual graphs for each document and then use GNN to learn the fine-grained word representations based on their local structure, which can also effectively produce embeddings for unseen words in the new document. Finally, the word nodes are aggregated as the document embedding. Extensive experiments on four benchmark datasets show that our method outperforms state-of-the-art text classification methods.

Neural Topic Modeling with Bidirectional Adversarial Training [Website][PDF]
Rui Wang, Xuemeng Hu, Deyu Zhou, Yulan He, Yuxuan Xiong, Chenchen Ye, and Haiyang Xu 5:00–6:00

Recent years have witnessed a surge of interests of using neural topic models for automatic topic extraction from text, since they avoid the complicated mathematical derivations for model inference as in traditional topic models such as Latent Dirichlet Allocation (LDA). However, these models either typically assume improper prior (e.g. Gaussian or Logistic Normal) over latent topic space or could not infer topic distribution for a given document. To address these limitations, we propose a neural topic modeling approach, called Bidirectional Adversarial Topic (BAT) model, which represents the first attempt of applying bidirectional adversarial training for neural topic modeling. The proposed BAT builds a two-way projection between the document-topic distribution and the document-word distribution. It uses a generator to capture the semantic patterns from texts and an encoder for topic inference. Furthermore, to incorporate word relatedness information, the Bidirectional Adversarial Topic model with Gaussian (Gaussian-BAT) is extended from BAT. To verify the effectiveness of BAT and Gaussian-BAT, three benchmark corpora are used in our experiments. The experimental results show that BAT and Gaussian-BAT obtain more coherent topics, outperforming several competitive baselines. Moreover, when performing text clustering based on the extracted topics, our models outperform all the baselines, with more significant improvements achieved by Gaussian-BAT where an increase of near 6% is observed in accuracy.

Text Classification with Negative Supervision [Website][PDF]
Sora Ohashi, Junya Takayama, Tomoyuki Kajiwara, Chenhui Chu, and Yuki Arase 5:00–6:00

Advanced pre-trained models for text representation have achieved state-of-the-art performance on various text classification tasks. However, the discrepancy between the semantic similarity of texts and labelling standards affects classifiers, i.e. leading to lower performance in cases where classifiers should assign different labels to semantically similar texts. To address this problem, we propose a simple multitask learning model that uses negative supervision. Specifically, our model encourages texts with different labels to have distinct representations. Comprehensive experiments show that our model outperforms the state-of-the-art pre-trained model on both single- and multi-label classifications, sentence and document classifications, and classifications in three different languages.

Session 1A: Machine Translation-1

Content Word Aware Neural Machine Translation [Website][PDF]
Kehai Chen, Rui Wang, Masao Utiyama, and Eiichiro Sumita 5:00–6:00

Neural machine translation (NMT) encodes the source sentence in a universal way to generate the target sentence word-by-word. However, NMT does not consider the importance of word in the sentence meaning, for example, some words (i.e., content words) express more important meaning than others (i.e., function words). To address this limitation, we first utilize word frequency information to distinguish between content and function words in a sentence, and then design a content word-aware NMT to improve translation performance. Empirical results on the WMT14 English-to-German, WMT14 English-to-French, and WMT17 Chinese-to-English translation tasks show that the proposed methods can significantly improve the performance of Transformer-based NMT.

Evaluating Explanation Methods for Neural Machine Translation [Website][PDF]
Jierui Li, Lemao Liu, Huayang Li, Guanlin Li, Guoping Huang, and Shuming Shi 5:00–6:00

Recently many efforts have been devoted to interpreting the black-box NMT models, but little progress has been made on metrics to evaluate explanation methods. Word Alignment Error Rate can be used as such a metric that matches human understanding, however, it can not measure explanation methods on those target words that are not aligned to any source word. This paper thereby makes an initial attempt to evaluate explanation methods from an alternative viewpoint. To this end, it proposes a principled metric based on fidelity in regard to the predictive behavior of the NMT model. As the exact computation for this metric is intractable, we employ an efficient approach as its approximation. On six standard translation tasks, we quantitatively evaluate several explanation methods in terms of the proposed metric and we reveal some valuable findings for these explanation methods in our experiments.

Jointly Masked Sequence-to-Sequence Model for Non-Autoregressive Neural Machine Translation [Website][PDF]
Junliang Guo, Linli Xu, and Enhong Chen 5:00–6:00

The masked language model has received remarkable attention due to its effectiveness on various natural language processing tasks. However, few works have adopted this technique in the sequence-to-sequence models. In this work, we introduce a jointly masked sequence-to-sequence model and explore its application on non-autoregressive neural machine translation~(NAT). Specifically, we first empirically study the functionalities of the encoder and the decoder in NAT models, and find that the encoder takes a more important role than the decoder regarding the translation quality. Therefore, we propose to train the encoder more rigorously by masking the encoder input while training. As for the decoder, we propose to train it based on the consecutive masking of the decoder input with an n-gram loss function to alleviate the problem of translating duplicate words. The two types of masks are applied to the model jointly at the training stage. We conduct experiments on five benchmark machine translation tasks, and our model can achieve 27.69/32.24 BLEU scores on WMT14 English-German/German-English tasks with $5+$ times speed up compared with an autoregressive model.

Learning Source Phrase Representations for Neural Machine Translation [Website][PDF]
Hongfei Xu, Josef van Genabith, Deyi Xiong, Qiuhui Liu, and Jingyi Zhang 5:00–6:00

The Transformer translation model (Vaswani et al., 2017) based on a multi-head attention mechanism can be computed effectively in parallel and has significantly pushed forward the performance of Neural Machine Translation (NMT). Though intuitively the attentional network can connect distant words via shorter network paths than RNNs, empirical analysis demonstrates that it still has difficulty in fully capturing long-distance dependencies (Tang et al., 2018). Considering that modeling phrases instead of words has significantly improved the Statistical Machine Translation (SMT) approach through the use of larger translation blocks ("phrases") and its reordering ability, modeling NMT at phrase level is an intuitive proposal to help the model capture long-distance relationships. In this paper, we first propose an attentive phrase representation generation mechanism which is able to generate phrase representations from corresponding token representations. In addition, we incorporate the generated phrase representations into the Transformer translation model to enhance its ability to capture long-distance relationships. In our experiments, we obtain significant improvements on the WMT 14 English-German and English-French tasks on top of the strong Transformer baseline, which shows the effectiveness of our approach. Our approach helps Transformer Base models perform at the level of Transformer Big models, and even significantly better for long sentences, but with substantially fewer parameters and training steps. The fact that phrase representations help even in the big setting further supports our conjecture that they make a valuable contribution to long-distance relations.

Lipschitz Constrained Parameter Initialization for Deep Transformers [Website][PDF]
Hongfei Xu, Qiuhui Liu, Josef van Genabith, Deyi Xiong, and Jingyi Zhang 5:00–6:00

The Transformer translation model employs residual connection and layer normalization to ease the optimization difficulties caused by its multi-layer encoder/decoder structure. Previous research shows that even with residual connection and layer normalization, deep Transformers still have difficulty in training, and particularly Transformer models with more than 12 encoder/decoder layers fail to converge. In this paper, we first empirically demonstrate that a simple modification made in the official implementation, which changes the computation order of residual connection and layer normalization, can significantly ease the optimization of deep Transformers. We then compare the subtle differences in computation order in considerable detail, and present a parameter initialization method that leverages the Lipschitz constraint on the initialization of Transformer parameters that effectively ensures training convergence. In contrast to findings in previous research we further demonstrate that with Lipschitz parameter initialization, deep Transformers with the original computation order can converge, and obtain significant BLEU improvements with up

to 24 layers. In contrast to previous research which focuses on deep encoders, our approach additionally enables Transformers to also benefit from deep decoders.

Location Attention for Extrapolation to Longer Sequences [Website][PDF]
Yann Dubois, Gautier Dagan, Dieuwke Hupkes, and Elia Bruni 5:00–6:00

Neural networks are surprisingly good at interpolating and perform remarkably well when the training set examples resemble those in the test set. However, they are often unable to extrapolate patterns beyond the seen data, even when the abstractions required for such patterns are simple. In this paper, we first review the notion of extrapolation, why it is important and how one could hope to tackle it. We then focus on a specific type of extrapolation which is especially useful for natural language processing: generalization to sequences that are longer than the training ones. We hypothesize that models with a separate content- and location-based attention are more likely to extrapolate than those with common attention mechanisms. We empirically support our claim for recurrent seq2seq models with our proposed attention on variants of the Lookup Table task. This sheds light on some striking failures of neural models for sequences and on possible methods to approaching such issues.

Multiscale Collaborative Deep Models for Neural Machine Translation [Website][PDF]
Xiangpeng Wei, Heng Yu, Yue Hu, Yue Zhang, Rongxiang Weng, and Weihua Luo 5:00–6:00

Recent evidence reveals that Neural Machine Translation (NMT) models with deeper neural networks can be more effective but are difficult to train. In this paper, we present a MultiScale Collaborative (MSC) framework to ease the training of NMT models that are substantially deeper than those used previously. We explicitly boost the gradient back-propagation from top to bottom levels by introducing a block-scale collaboration mechanism into deep NMT models. Then, instead of forcing the whole encoder stack directly learns a desired representation, we let each encoder block learns a fine-grained representation and enhance it by encoding spatial dependencies using a context-scale collaboration. We provide empirical evidence showing that the MSC nets are easy to optimize and can obtain improvements of translation quality from considerably increased depth. On IWSLT translation tasks with three translation directions, our extremely deep models (with 72-layer encoders) surpass strong baselines by +2.2∼+3.1 BLEU points. In addition, our deep MSC achieves a BLEU score of 30.56 on WMT14 English-to-German task that significantly outperforms state-of-the-art deep NMT models. We have included the source code in supplementary materials.

Norm-Based Curriculum Learning for Neural Machine Translation [Website][PDF]
Xuebo Liu, Houtim Lai, Derek F. Wong, and Lidia S. Chao 5:00–6:00

A neural machine translation (NMT) system is expensive to train, especially with high-resource settings. As the NMT architectures become deeper and wider, this issue gets worse and worse. In this paper, we aim to improve the efficiency of training an NMT by introducing a novel norm-based curriculum learning method. We use the norm (aka length or module) of a word embedding as a measure of 1) the difficulty of the sentence, 2) the competence of the model, and 3) the weight of the sentence. The norm-based sentence difficulty takes the advantages of both linguistically motivated and model-based sentence difficulties. It is easy to determine and contains learning-dependent features. The norm-based model competence makes NMT learn the curriculum in a fully automated way, while the norm-based sentence weight further enhances the learning of the vector representation of the NMT. Experimental results for the WMT'14 English-German and WMT'17 Chinese-English translation tasks demonstrate that the proposed method outperforms strong baselines in terms of BLEU score (+1.17/+1.56) and training speedup (2.22x/3.33x).

Opportunistic Decoding with Timely Correction for Simultaneous Translation [Website][PDF]
Renjie Zheng, Mingbo Ma, Baigong Zheng, Kaibo Liu, and Liang Huang 5:00–6:00

Simultaneous translation has many important application scenarios and attracts much attention from both academia and industry recently. Most existing frameworks, however, have difficulties in balancing between the translation quality and latency, i.e., the decoding policy is usually either too aggressive or too conservative. We propose an opportunistic decoding technique with timely correction ability, which always (over-)generates a certain mount of extra words at each step to keep the audience on track with the latest information. At the same time, it also corrects, in a timely fashion, the mistakes in the former overgenerated words when observing more source context to ensure high translation quality. Experiments show our technique achieves substantial reduction in latency and up to +3.1 increase in BLEU, with revision rate under 8% in Chinese-to-English and English-to-Chinese translation.

Session 1A: Student Research Workshop

Adaptive Transformers for Learning Multimodal Representations [Website][PDF]
Prajjwal Bhargava 5:00–6:00

The usage of transformers has grown from learning about language semantics to forming meaningful visiolinguistic representations. These architectures are often over-parametrized, requiring large amounts of computation. In this work, we extend adaptive approaches to learn more about model interpretability and computational efficiency. Specifically, we study attention spans, sparse, and structured dropout methods to help understand how their attention mechanism extends for vision and language tasks. We further show that these approaches can help us learn more about how the network perceives the complexity of input sequences, sparsity preferences for different modalities, and other related phenomena.

Story-level Text Style Transfer: A Proposal [Website][PDF]
Yusu Qian 5:00–6:00

Text style transfer aims to change the style of the input text to the target style while preserving the content to some extent. Previous works on this task are on the sentence level. We aim to work on story-level text style transfer to generate stories that preserve the plot of the input story while exhibiting a strong target style. The challenge in this task compared to previous work is that the structure of the input story, consisting of named entities and their relations with each other, needs to be preserved, and that the generated story needs to be consistent after adding flavors. We plan to explore three methods including the BERT-based method, the Story Realization method, and the Graph-based method.

Unsupervised Paraphasia Classification in Aphasic Speech [Website][PDF]
Sharan Pai, Nikhil Sachdeva, Prince Sachdeva, and Rajiv Ratn Shah 5:00–6:00

Aphasia is a speech and language disorder which results from brain damage, often characterized by word retrieval deficit (anomia) resulting in naming errors (paraphasia). Automatic paraphasia detection has many benefits for both treatment and diagnosis of Aphasia and its type. But supervised learning methods cant be properly utilized as there is a lack of aphasic speech data. In this paper, we describe our novel unsupervised method which can be implemented without the need for labeled paraphasia data. Our evaluations show that our method outperforms previous work based on supervised learning and transfer learning approaches for English. We demonstrate the utility of our method as an essential first step in developing augmentative and alternative communication (AAC) devices for patients suffering from aphasia in any language.

HGCN4MeSH: Hybrid Graph Convolution Network for MeSH Indexing [Website][PDF]
Miaomiao Yu, Yujiu Yang, and Chenhui Li 5:00–6:00

Recently deep learning has been used in Medical subject headings (MeSH) indexing to reduce the time and monetary cost by manual annotation, including DeepMeSH, TextCNN, etc. However, these models still suffer from failing to capture the complex correlations between MeSH terms. To this end, we introduce Graph Convolution Network (GCN) to learn the relationship between these terms, and present a novel Hybrid Graph Convolution Net for MeSH index (HGCN4MeSH). Basically, we utilize two BiGRUs to learn the embedding representation of the abstract and the title of the MeSH index text respectively. At the same time, we establish the adjacency matrix of MeSH terms based on the co-occurrence relationships in Corpus, which is easy to apply for GCN representation learning. On the basis of learning the mixed representation, the prediction problem of the MeSH index keywords is transformed into an extreme multi-label classification problem after the attention layer operation. Experimental results on two datasets show that HGCN4MeSH is competitive compared with the state-of-the-art methods.

Session 1A: Theory and Formalism in NLP (Linguistic and Mathematical)-1

A Formal Hierarchy of RNN Architectures [Website][PDF]
William Merrill, Gail Weiss, Yoav Goldberg, Roy Schwartz, Noah A. Smith, and Eran Yahav 5:00–6:00

We develop a formal hierarchy of the expressive capacity of RNN architectures. The hierarchy is based on two formal properties: space complexity, which measures the RNN's memory, and rational recurrence, defined as whether the recurrent update can be described by a weighted finite-state machine. We place several RNN variants within this hierarchy. For example, we prove the LSTM is not rational, which formally separates it from the related QRNN (Bradbury et al., 2016). We also show how these models' expressive capacity is expanded by stacking multiple layers or composing them with different pooling functions. Our results build on the theory of "saturated" RNNs (Merrill, 2019). While formally extending these findings to unsaturated RNNs is left to future work, we hypothesize that the practical learnable capacity of unsaturated RNNs obeys a similar hierarchy. We provide empirical results to support this conjecture. Experimental findings from training unsaturated networks on formal languages support this conjecture.

A Three-Parameter Rank-Frequency Relation in Natural Languages [Website][PDF]
Chenchen Ding, Masao Utiyama, and Eiichiro Sumita 5:00–6:00

We present that, the rank-frequency relation in textual data follows $f \propto r^{-\alpha}(r+\gamma)^{-\beta}$, where f is the token frequency and r is the rank by frequency, with (α, β, γ) as parameters. The formulation is derived based on the empirical observation that $d^2(x+y)/dx^2$ is a typical impulse function, where $(x, y) = (\log r, \log f)$. The formulation is the power law when $\beta = 0$ and the Zipf–Mandelbrot law when $\alpha = 0$. We illustrate that α is related to the analytic features of syntax and $\beta+\gamma$ to those of morphology in natural languages from an investigation of multilingual corpora.

Dice Loss for Data-imbalanced NLP Tasks [Website][PDF]
Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li 5:00–6:00

Many NLP tasks such as tagging and machine reading comprehension are faced with the severe data imbalance issue: negative examples significantly outnumber positive examples, and the huge number of easy-negative examples overwhelms the training. The most commonly used cross entropy (CE) criteria is actually an accuracy-oriented objective, and thus creates a discrepancy between training and test: at training time, each training instance contributes equally to the objective function, while at test time F1 score concerns more about positive examples. In this paper, we propose to use dice loss in replacement of the standard cross-entropy objective for data-imbalanced NLP tasks. Dice loss is based on the Sørensen—Dice coefficient or Tversky index , which attaches similar importance to false positives and false negatives, and is more immune to the data-imbalance issue. To further alleviate the dominating influence from easy-negative examples in training, we propose to associate training examples with dynamically adjusted weights to deemphasize easy-negative examples. Theoretical analysis shows that this strategy narrows down the gap between the F1 score in evaluation and the dice loss in training. With the proposed training objective, we observe significant performance boost on a wide range of data imbalanced NLP tasks. Notably, we are able to achieve SOTA results on CTB5, CTB6 and UD1.4 for the part of speech tagging task; SOTA results on CoNLL03, OntoNotes5.0, MSRA and OntoNotes4.0 for the named entity recognition task; along with competitive results on the tasks of machine reading comprehension and paraphrase identification.

Emergence of Syntax Needs Minimal Supervision [Website][PDF]
Raphaël Bailly and Kata Gábor 5:00–6:00

This paper is a theoretical contribution to the debate on the learnability of syntax from a corpus without explicit syntax-specific guidance. Our approach originates in the observable structure of a corpus, which we use to define and isolate grammaticality (syntactic information) and meaning/pragmatics information. We describe the formal characteristics of an autonomous syntax and show that it becomes possible to search for syntax-based lexical categories with a simple optimization process, without any prior hypothesis on the form of the model.

Language Models as an Alternative Evaluator of Word Order Hypotheses: A Case Study in Japanese [Website][PDF]
Tatsuki Kuribayashi, Takumi Ito, Jun Suzuki, and Kentaro Inui 5:00–6:00

We examine a methodology using neural language models (LMs) for analyzing the word order of language. This LM-based method has the potential to overcome the difficulties existing methods face, such as the propagation of preprocessor errors in count-based methods. In this study, we explore whether the LM-based method is valid for analyzing the word order. As a case study, this study focuses on Japanese due to its complex and flexible word order. To validate the LM-based method, we test (i) parallels between LMs and human word order preference, and (ii) consistency of the results obtained using the LM-based method with previous linguistic studies. Through our experiments, we tentatively conclude that LMs display sufficient word order knowledge for usage as an analysis tool. Finally, using the LM-based method, we demonstrate the relationship between the canonical word order and topicalization, which had yet to be analyzed by large-scale experiments.

[TACL] Theoretical Limitations of Self-Attention in Neural Sequence Models [Website][PDF]
Michael Hahn 5:00–6:00

Transformers are emerging as the new workhorse of NLP, showing great success across tasks. Unlike LSTMs, transformers process input sequences entirely through self-attention. Previous work has suggested that the computational capabilities of self-attention to process hierarchical structures are limited. In this work, we mathematically investigate the computational power of self-attention to model formal languages. Across both soft and hard attention, we show strong theoretical limitations of the computational abilities of self-attention, finding that it cannot model periodic

finite-state languages, nor hierarchical structure, unless the number of layers or heads increases with input length. These limitations seem surprising given the practical success of self-attention and the prominent role assigned to hierarchical structure in linguistics, suggesting that natural language can be approximated well with models that are too weak for the formal languages typically assumed in theoretical linguistics.

Demo Session 1B

Time: *5:45–6:30*

TextBrewer: An Open-Source Knowledge Distillation Toolkit for Natural Language Processing [Website] [PDF]

Ziqing Yang, Yiming Cui, Zhipeng Chen, Wanxiang Che, Ting Liu, Shijin Wang, and Guoping Hu

In this paper, we introduce TextBrewer, an open-source knowledge distillation toolkit designed for natural language processing. It works with different neural network models and supports various kinds of supervised learning tasks, such as text classification, reading comprehension, sequence labeling. TextBrewer provides a simple and uniform workflow that enables quick setting up of distillation experiments with highly flexible configurations. It offers a set of predefined distillation methods and can be extended with custom code. As a case study, we use TextBrewer to distill BERT on several typical NLP tasks. With simple configurations, we achieve results that are comparable with or even higher than the public distilled BERT models with similar numbers of parameters.

Session 1B Overview – Monday, July 6, 2020 6:00–7:00

Track A *Computational Social Science and Social Media-1* Abstracts	GCAN: Graph-aware Co-Attention Networks for Explainable Fake News Detection on Social Media *Lu and Li* [Website][PDF]	Integrating Semantic and Structural Information with Graph Convolutional Network for Controversy Detection *Zhong, Cao, Sheng, Guo, and Wang* [Website][PDF]	Predicting the Topical Stance and Political Leaning of Media using Tweets *Stefanov, Darwish, Atanasov, and Nakov* [Website][PDF]	Simple, Interpretable and Stable Method for Detecting Words with Usage Change across Corpora *Gonen, Jawahar, Seddah, and Goldberg* [Website][PDF]	
Track B *Dialogue and Interactive Systems-2* Abstracts	CDL: Curriculum Dual Learning for Emotion-Controllable Response Generation *Shen and Feng* [Website][PDF]	[TACL] Cross-WOZ: A Large-Scale Chinese Cross-Domain Task-Oriented Dialogue Dataset *Zhu, Huang, Zhang, Zhu, and Huang* [Website][PDF]	Efficient Dialogue State Tracking by Selectively Overwriting Memory *Kim, Yang, Kim, and Lee* [Website][PDF]	End-to-End Neural Pipeline for Goal-Oriented Dialogue Systems using GPT-2 *Ham, Lee, Jang, and Kim* [Website][PDF]	Evaluating Dialogue Generation Systems via Response Selection *Sato, Akama, Ouchi, Suzuki, and Inui* [Website][PDF]
	Gated Convolutional Bidirectional Attention-based Model for Off-topic Spoken Response Detection *Zha, Li, and Lin* [Website][PDF]	Learning Low-Resource End-To-End Goal-Oriented Dialog for Fast and Reliable System Deployment *Dai, Li, Tang, Li, Sun, and Zhu* [Website][PDF]	Learning to Tag OOV Tokens by Integrating Contextual Representation and Background Knowledge *He, Yan, and XU* [Website][PDF]	Multi-Agent Task-Oriented Dialog Policy Learning with Role-Aware Reward Decomposition *Takanobu, Liang, and Huang* [Website][PDF]	Paraphrase Augmented Task-Oriented Dialog Generation *Gao, Zhang, Ou, and Yu* [Website][PDF]
	Response-Anticipated Memory for On-Demand Knowledge Integration in Response Generation *Tian, Bi, Lee, Xue, SONG, Liu, and Zhang* [Website][PDF]	Semi-Supervised Dialogue Policy Learning via Stochastic Reward Estimation *Huang, Qi, Sun, and Zhang* [Website][PDF]	Towards Unsupervised Language Understanding and Generation by Joint Dual Learning *Su, Huang, and Chen* [Website][PDF]	USR: An Unsupervised and Reference Free Evaluation Metric for Dialog Generation *Mehri and Eskenazi* [Website][PDF]	
Track C *Generation-2* Abstracts	Explicit Semantic Decomposition for Definition Generation *Li, Bao, Huang, Dai, and CHEN* [Website][PDF]	Improved Natural Language Generation via Loss Truncation *Kang and Hashimoto* [Website][PDF]	Line Graph Enhanced AMR-to-Text Generation with Mix-Order Graph Attention Networks *Zhao, Chen, Chen, Cao, Zhu, and Yu* [Website][PDF]	Rigid Formats Controlled Text Generation *Li, Zhang, Liu, and Shi* [Website][PDF]	Syn-QG: Syntactic and Shallow Semantic Rules for Question Generation *Dhole and Manning* [Website][PDF]
Track D *Information Retrieval and Text Mining-2* Abstracts	An Online Semantic-enhanced Dirichlet Model for Short Text Stream Clustering *Kumar, Shao, Uddin, and Ali* [Website][PDF]	Generative Semantic Hashing Enhanced via Boltzmann Machines *Zheng, Su, Shen, and Chen* [Website][PDF]	Interactive Construction of User-Centric Dictionary for Text Analytics *Kohita, Yoshida, Kanayama, and Nasukawa* [Website][PDF]	Tree-Structured Neural Topic Model *Isonuma, Mori, Bollegala, and Sakata* [Website][PDF]	Unsupervised FAQ Retrieval with Question Generation and BERT *Mass, Carmeli, Roitman, and Konopnicki* [Website][PDF]

Track E *NLP Applications-1* Abstracts	"The Boating Store Had Its Best Sail Ever": Pronunciation-attentive Contextualized Pun Recognition *Zhou, Jiang, Zhao, Chang, and Wang* [Website][PDF]	Fast and Accurate Deep Bidirectional Language Representations for Unsupervised Learning *Shin, Lee, Yoon, and Jung* [Website][PDF]	Fine-grained Interest Matching for Neural News Recommendation *Wang, Wu, Liu, and Xie* [Website][PDF]	Interpretable Operational Risk Classification with Semi-Supervised Variational Autoencoder *Zhou, Zhang, and Yang* [Website][PDF]	Interpreting Twitter User Geolocation *Zhong, Wang, Zhou, Trajcevski, Zhang, and Yang* [Website][PDF]
	Modeling Code-Switch Languages Using Bilingual Parallel Corpus *Lee and Li* [Website][PDF]	SpellGCN: Incorporating Phonological and Visual Similarities into Language Models for Chinese Spelling Check *Cheng, Xu, Chen, Jiang, Wang, Wang, Chu, and Qi* [Website][PDF]	Spelling Error Correction with Soft-Masked BERT *Zhang, Huang, Liu, and Li* [Website][PDF]		
Track F *Question Answering-1* Abstracts	A Frame-based Sentence Representation for Machine Reading Comprehension *Guo, Li, Tan, Li, Guan, Zhao, and Zhang* [Website][PDF]	A Methodology for Creating Question Answering Corpora Using Inverse Data Annotation *Deriu, Mlynchyk, Schläpfer, Rodrigo, Grünigen, Kaiser, Stockinger, Agirre, and Cieliebak* [Website][PDF]	Contextualized Sparse Representations for Real-Time Open-Domain Question Answering *Lee, Seo, Hajishirzi, and Kang* [Website][PDF]	Dynamic Sampling Strategies for Multi-Task Reading Comprehension *Gottumukkala, Dua, Singh, and Gardner* [Website][PDF]	Enhancing Answer Boundary Detection for Multilingual Machine Reading Comprehension *Yuan, Shou, Bai, Gong, Liang, Duan, Fu, and Jiang* [Website][PDF]
	Explicit Memory Tracker with Coarse-to-Fine Reasoning for Conversational Machine Reading *Gao, Wu, Joty, Xiong, Socher, King, Lyu, and Hoi* [Website][PDF]	Injecting Numerical Reasoning Skills into Language Models *Geva, Gupta, and Berant* [Website][PDF]	Learning to Identify Follow-Up Questions in Conversational Question Answering *Kundu, Lin, and Ng* [Website][PDF]	Query Graph Generation for Answering Multi-hop Complex Questions from Knowledge Bases *Lan and Jiang* [Website][PDF]	
Track G *Resources and Evaluation-1* Abstracts	A Diverse Corpus for Evaluating and Developing English Math Word Problem Solvers *Miao, Liang, and Su* [Website][PDF]	Improving Image Captioning Evaluation by Considering Inter References Variance *Yi, Deng, and Hu* [Website][PDF]	Revisiting the Context Window for Cross-lingual Word Embeddings *Ri and Tsuruoka* [Website][PDF]		
Track H *Lexical-1* Abstracts	Moving Down the Long Tail of Word Sense Disambiguation with Gloss Informed Bi-encoders *Blevins and Zettlemoyer* [Website][PDF]				

Track I *Student Research Workshop* Abstracts	Grammatical Error Correction Using Pseudo Learner Corpus Considering Learner's Error Tendency *Takahashi, Katsumata, and Komachi* [Website] [PDF]	Research on Task Discovery for Transfer Learning in Deep Neural Networks *Akdemir* [Website] [PDF]	RPD: A Distance Function Between Word Embeddings *Zhou, Huang, and Zheng* [Website] [PDF]	Reflection-based Word Attribute Transfer *Ishibashi, Sudoh, Yoshino, and Nakamura* [Website] [PDF]	

Session 1B Details

Session 1B: Computational Social Science and Social Media-1

GCAN: Graph-aware Co-Attention Networks for Explainable Fake News Detection on Social Media
[Website][PDF]
Yi-Ju Lu and Cheng-Te Li 6:00–7:00

This paper solves the fake news detection problem under a more realistic scenario on social media. Given the source short-text tweet and the corresponding sequence of retweet users without text comments, we aim at predicting whether the source tweet is fake or not, and generating explanation by highlighting the evidences on suspicious retweeters and the words they concern. We develop a novel neural network-based model, Graph-aware Co-Attention Networks (GCAN), to achieve the goal. Extensive experiments conducted on real tweet datasets exhibit that GCAN can significantly outperform state-of-the-art methods by 16% in accuracy on average. In addition, the case studies also show that GCAN can produce reasonable explanations.

Integrating Semantic and Structural Information with Graph Convolutional Network for Controversy Detection [Website][PDF]
Lei Zhong, Juan Cao, Qiang Sheng, Junbo Guo, and Ziang Wang 6:00–7:00

Identifying controversial posts on social media is a fundamental task for mining public sentiment, assessing the influence of events, and alleviating the polarized views. However, existing methods fail to 1) effectively incorporate the semantic information from content-related posts; 2) preserve the structural information for reply relationship modeling; 3) properly handle posts from topics dissimilar to those in the training set. To overcome the first two limitations, we propose Topic-Post-Comment Graph Convolutional Network (TPC-GCN), which integrates the information from the graph structure and content of topics, posts, and comments for post-level controversy detection. As to the third limitation, we extend our model to Disentangled TPC-GCN (DTPC-GCN), to disentangle topic-related and topic-unrelated features and then fuse dynamically. Extensive experiments on two real-world datasets demonstrate that our models outperform existing methods. Analysis of the results and cases proves that our models can integrate both semantic and structural information with significant generalizability.

Predicting the Topical Stance and Political Leaning of Media using Tweets [Website][PDF]
Peter Stefanov, Kareem Darwish, Atanas Atanasov, and Preslav Nakov 6:00–7:00

Discovering the stances of media outlets and influential people on current, debatable topics is important for social statisticians and policy makers. Many supervised solutions exist for determining viewpoints, but manually annotating training data is costly. In this paper, we propose a cascaded method that uses unsupervised learning to ascertain the stance of Twitter users with respect to a polarizing topic by leveraging their retweet behavior; then, it uses supervised learning based on user labels to characterize both the general political leaning of online media and of popular Twitter users, as well as their stance with respect to the target polarizing topic. We evaluate the model by comparing its predictions to gold labels from the Media Bias/Fact Check website, achieving 82.6% accuracy.

Simple, Interpretable and Stable Method for Detecting Words with Usage Change across Corpora
[Website][PDF]
Hila Gonen, Ganesh Jawahar, Djamé Seddah, and Yoav Goldberg 6:00–7:00

The problem of comparing two bodies of text and searching for words that differ in their usage between them arises often in digital humanities and computational social science. This is commonly approached by training word embeddings on each corpus, aligning the vector spaces, and looking for words whose cosine distance in the aligned space is large. However, these methods often require extensive filtering of the vocabulary to perform well, and - as we show in this work - result in unstable, and hence less reliable, results. We propose an alternative approach that does not use vector space alignment, and instead considers the neighbors of each word. The method is simple, interpretable and stable. We demonstrate its effectiveness in 9 different setups, considering different corpus splitting criteria (age, gender and profession of tweet authors, time of tweet) and different languages (English, French and Hebrew).

Session 1B: Dialogue and Interactive Systems-2

CDL: Curriculum Dual Learning for Emotion-Controllable Response Generation [Website][PDF]
Lei Shen and Yang Feng 6:00–7:00

Emotion-controllable response generation is an attractive and valuable task that aims to make open-domain conversations more empathetic and engaging. Existing methods mainly enhance the emotion expression by adding regularization terms to standard cross-entropy loss and thus influence the training process. However, due to the lack of further consideration of content consistency, the common problem of response generation tasks, safe response, is intensified. Besides, query emotions that can help model the relationship between query and response are simply ignored in previous models, which would further hurt the coherence. To alleviate these problems, we propose a novel framework named Curriculum Dual Learning (CDL) which extends the emotion-controllable response generation to a dual task to generate emotional responses and emotional queries alternatively. CDL utilizes two rewards focusing on emotion and content to improve the duality. Additionally, it applies curriculum learning to gradually generate high-quality responses based on the difficulties of expressing various emotions. Experimental results show that CDL significantly outperforms the baselines in terms of coherence, diversity, and relation to emotion factors.

[TACL] CrossWOZ: A Large-Scale Chinese Cross-Domain Task-Oriented Dialogue Dataset [Website][PDF]
Qi Zhu, Kaili Huang, Zheng Zhang, Xiaoyan Zhu, and Minlie Huang 6:00–7:00

To advance multi-domain (cross-domain) dialogue modeling as well as alleviate the shortage of Chinese task-oriented datasets, we propose CrossWOZ, the first large-scale Chinese Cross-Domain Wizard-of-Oz task-oriented dataset. It contains 6K dialogue sessions and 102K utterances for 5 domains, including hotel, restaurant, attraction, metro, and taxi. Moreover, the corpus contains rich annotation of dialogue states and dialogue acts at both user and system sides. About 60% of the dialogues have cross-domain user goals that favor inter-domain dependency and encourage natural transition across domains in conversation. We also provide a user simulator and several benchmark models for pipelined task-oriented dialogue systems, which will facilitate researchers to compare and evaluate their models on this corpus. The large size and rich annotation of CrossWOZ make it suitable to investigate a variety of tasks in cross-domain dialogue modeling, such as dialogue state tracking, policy learning, user simulation, etc.

Efficient Dialogue State Tracking by Selectively Overwriting Memory [Website][PDF]
Sungdong Kim, Sohee Yang, Gyuwan Kim, and Sang-Woo Lee 6:00–7:00

Recent works in dialogue state tracking (DST) focus on an open vocabulary-based setting to resolve scalability and generalization issues of the predefined ontology-based approaches. However, they are inefficient in that they predict the dialogue state at every turn from scratch. Here, we consider dialogue state as an explicit fixed-sized memory and propose a selectively overwriting mechanism for more efficient DST. This mechanism consists of two steps: (1) predicting state operation on each of the memory slots, and (2) overwriting the memory with new values, of which only a few are generated according to the predicted state operations. Our method decomposes DST into two sub-tasks and guides the decoder to focus only on one of the tasks, thus reducing the burden of the decoder. This enhances the effectiveness of training and DST performance. Our SOM-DST (Selectively Overwriting Memory for Dialogue State Tracking) model achieves state-of-the-art joint goal accuracy with 51.72% in MultiWOZ 2.0 and 53.01% in MultiWOZ 2.1 in an open vocabulary-based DST setting. In addition, we analyze the accuracy gaps between the current and the ground truth-given situations and suggest that it is a promising direction to improve state operation prediction to boost the DST performance.

End-to-End Neural Pipeline for Goal-Oriented Dialogue Systems using GPT-2 [Website][PDF]
Donghoon Ham, Jeong-Gwan Lee, Youngsoo Jang, and Kee-Eung Kim 6:00–7:00

The goal-oriented dialogue system needs to be optimized for tracking the dialogue flow and carrying out an effective conversation under various situations to meet the user goal. The traditional approach to build such a dialogue system is to take a pipelined modular architecture, where its modules are optimized individually. However, such an optimization scheme does not necessarily yield the overall performance improvement of the whole system. On the other hand, end-to-end dialogue systems with monolithic neural architecture are often trained only with input-output utterances, without taking into account the entire annotations available in the corpus. This scheme makes it difficult for goal-oriented dialogues where the system needs to integrate with external systems or to provide interpretable information about why the system generated a particular response. In this paper, we present an end-to-end neural architecture for dialogue systems that addresses both challenges above. In the human evaluation, our dialogue system achieved the success rate of 68.32%, the language understanding score of 4.149, and the response appropriateness score of 4.287, which ranked the system at the top position in the end-to-end multi-domain dialogue system task in the 8th dialogue systems technology challenge (DSTC8).

Evaluating Dialogue Generation Systems via Response Selection [Website][PDF]
Shiki Sato, Reina Akama, Hiroki Ouchi, Jun Suzuki, and Kentaro Inui 6:00–7:00

Existing automatic evaluation metrics for open-domain dialogue response generation systems correlate poorly with human evaluation. We focus on evaluating response generation systems via response selection. To evaluate systems properly via response selection, we propose a method to construct response selection test sets with well-chosen false candidates. Specifically, we propose to construct test sets filtering out some types of false candidates: (i) those unrelated to the ground-truth response and (ii) those acceptable as appropriate responses. Through experiments, we demonstrate that evaluating systems via response selection with the test set developed by our method correlates more strongly with human evaluation, compared with widely used automatic evaluation metrics such as BLEU.

Gated Convolutional Bidirectional Attention-based Model for Off-topic Spoken Response Detection
[Website][PDF]
Yefei Zha, Ruobing Li, and Hui Lin 6:00–7:00

Off-topic spoken response detection, the task aiming at predicting whether a response is off-topic for the corresponding prompt, is important for an automated speaking assessment system. In many real-world educational applications, off-topic spoken response detectors are required to achieve high recall for off-topic responses not only on seen prompts but also on prompts that are unseen during training. In this paper, we propose a novel approach for off-topic spoken response detection with high off-topic recall on both seen and unseen prompts. We introduce a new model, Gated Convolutional Bidirectional Attention-based Model (GCBiA), which applies bi-attention mechanism and convolutions to extract topic words of prompts and key-phrases of responses, and introduces gated unit and residual connections between major layers to better represent the relevance of responses and prompts. Moreover, a new negative sampling method is proposed to augment training data. Experiment results demonstrate that our novel approach can achieve significant improvements in detecting off-topic responses with extremely high on-topic recall, for both seen and unseen prompts.

Learning Low-Resource End-To-End Goal-Oriented Dialog for Fast and Reliable System Deployment
[Website][PDF]
Yinpei Dai, Hangyu Li, Chengguang Tang, Yongbin Li, Jian Sun, and Xiaodan Zhu 6:00–7:00

Existing end-to-end dialog systems perform less effectively when data is scarce. To obtain an acceptable success in real-life online services with only a handful of training examples, both fast adaptability and reliable performance are highly desirable for dialog systems. In this paper, we propose the Meta-Dialog System (MDS), which combines the advantages of both meta-learning approaches and human-machine collaboration. We evaluate our methods on a new extended-bAbI dataset and a transformed MultiWOZ dataset for low-resource goal-oriented dialog learning. Experimental results show that MDS significantly outperforms non-meta-learning baselines and can achieve more than 90% per-turn accuracies with only 10 dialogs on the extended-bAbI dataset.

Learning to Tag OOV Tokens by Integrating Contextual Representation and Background Knowledge
[Website][PDF]
Keqing He, Yuanmeng Yan, and Weiran XU 6:00–7:00

Neural-based context-aware models for slot tagging have achieved state-of-the-art performance. However, the presence of OOV(out-of-vocab) words significantly degrades the performance of neural-based models, especially in a few-shot scenario. In this paper, we propose a novel knowledge-enhanced slot tagging model to integrate contextual representation of input text and the large-scale lexical background knowledge. Besides, we use multi-level graph attention to explicitly model lexical relations. The experiments show that our proposed knowledge integration mechanism achieves consistent improvements across settings with different sizes of training data on two public benchmark datasets.

Multi-Agent Task-Oriented Dialog Policy Learning with Role-Aware Reward Decomposition [Website][PDF]
Ryuichi Takanobu, Runze Liang, and Minlie Huang 6:00–7:00

Many studies have applied reinforcement learning to train a dialog policy and show great promise these years. One common approach is to employ a user simulator to obtain a large number of simulated user experiences for reinforcement learning algorithms. However, modeling a realistic user simulator is challenging. A rule-based simulator requires heavy domain expertise for complex tasks, and a data-driven simulator requires considerable data and it is even unclear how to evaluate a simulator. To avoid explicitly building a user simulator beforehand, we propose Multi-Agent Dialog Policy Learning, which regards both the system and the user as the dialog agents. Two agents interact with each other and are jointly learned simultaneously. The method uses the actor-critic framework to facilitate pre-training and improve scalability. We also propose Hybrid Value Network for the role-aware reward decomposition to integrate role-specific domain knowledge of each agent in the task-oriented dialog. Results show that our method can successfully build a system policy and a user policy simultaneously, and two agents can achieve a high task success rate through conversational interaction.

Paraphrase Augmented Task-Oriented Dialog Generation [Website][PDF]
Silin Gao, Yichi Zhang, Zhijian Ou, and Zhou Yu 6:00–7:00

Neural generative models have achieved promising performance on dialog generation tasks if given a huge data set. However, the lack of high-quality dialog data and the expensive data annotation process greatly limit their application in real world settings. We propose a paraphrase augmented response generation (PARG) framework that jointly trains a paraphrase model and a response generation model to improve the dialog generation performance. We also design a method to automatically construct paraphrase training data set based on dialog state and dialog act labels. PARG is applicable to various dialog generation models, such as TSCP (Lei et al., 2018) and DAMD (Zhang et al., 2019). Experimental results show that the proposed framework improves these state-of-the-art dialog models further on CamRest676 and MultiWOZ. PARG also outperforms other data augmentation methods significantly in dialog generation tasks, especially under low resource settings.

Response-Anticipated Memory for On-Demand Knowledge Integration in Response Generation
[Website][PDF]
Zhiliang Tian, Wei Bi, Dongkyu Lee, Lanqing Xue, YIPING SONG, Xiaojiang Liu, and Nevin L. Zhang
6:00–7:00

Neural conversation models are known to generate appropriate but non-informative responses in general. A sce-

nario where informativeness can be significantly enhanced is Conversing by Reading (CbR), where conversations take place with respect to a given external document. In previous work, the external document is utilized by (1) creating a context-aware document memory that integrates information from the document and the conversational context, and then (2) generating responses referring to the memory. In this paper, we propose to create the document memory with some anticipated responses in mind. This is achieved using a teacher-student framework. The teacher is given the external document, the context, and the ground-truth response, and learns how to build a response-aware document memory from three sources of information. The student learns to construct a response-anticipated document memory from the first two sources, and teacher's insight on memory creation. Empirical results show that our model outperforms the previous state-of-the-art for the CbR task.

Semi-Supervised Dialogue Policy Learning via Stochastic Reward Estimation [Website][PDF]
Xinting Huang, Jianzhong Qi, Yu Sun, and Rui Zhang 6:00–7:00

Dialogue policy optimization often obtains feedback until task completion in task-oriented dialogue systems. This is insufficient for training intermediate dialogue turns since supervision signals (or rewards) are only provided at the end of dialogues. To address this issue, reward learning has been introduced to learn from state-action pairs of an optimal policy to provide turn-by-turn rewards. This approach requires complete state-action annotations of human-to-human dialogues (i.e., expert demonstrations), which is labor intensive. To overcome this limitation, we propose a novel reward learning approach for semi-supervised policy learning. The proposed approach learns a dynamics model as the reward function which models dialogue progress (i.e., state-action sequences) based on expert demonstrations, either with or without annotations. The dynamics model computes rewards by predicting whether the dialogue progress is consistent with expert demonstrations. We further propose to learn action embeddings for a better generalization of the reward function. The proposed approach outperforms competitive policy learning baselines on MultiWOZ, a benchmark multi-domain dataset.

Towards Unsupervised Language Understanding and Generation by Joint Dual Learning [Website][PDF]
Shang-Yu Su, Chao-Wei Huang, and Yun-Nung Chen 6:00–7:00

In modular dialogue systems, natural language understanding (NLU) and natural language generation (NLG) are two critical components, where NLU extracts the semantics from the given texts and NLG is to construct corresponding natural language sentences based on the input semantic representations. However, the dual property between understanding and generation has been rarely explored. The prior work is the first attempt that utilized the duality between NLU and NLG to improve the performance via a dual supervised learning framework. However, the prior work still learned both components in a supervised manner; instead, this paper introduces a general learning framework to effectively exploit such duality, providing flexibility of incorporating both supervised and unsupervised learning algorithms to train language understanding and generation models in a joint fashion. The benchmark experiments demonstrate that the proposed approach is capable of boosting the performance of both NLU and NLG. The source code is available at: https://github.com/MiuLab/DuaLUG.

USR: An Unsupervised and Reference Free Evaluation Metric for Dialog Generation [Website][PDF]
Shikib Mehri and Maxine Eskenazi 6:00–7:00

The lack of meaningful automatic evaluation metrics for dialog has impeded open-domain dialog research. Standard language generation metrics have been shown to be ineffective for evaluating dialog models. To this end, this paper presents USR, an UnSupervised and Reference-free evaluation metric for dialog. USR is a reference-free metric that trains unsupervised models to measure several desirable qualities of dialog. USR is shown to strongly correlate with human judgment on both Topical-Chat (turn-level: 0.42, system-level: 1.0) and PersonaChat (turn-level: 0.48 and system-level: 1.0). USR additionally produces interpretable measures for several desirable properties of dialog.

Session 1B: Generation-2

Explicit Semantic Decomposition for Definition Generation [Website][PDF]
Jiahuan Li, Yu Bao, Shujian Huang, Xinyu Dai, and Jiajun CHEN 6:00–7:00

Definition generation, which aims to automatically generate dictionary definitions for words, has recently been proposed to assist the construction of dictionaries and help people understand unfamiliar texts. However, previous works hardly consider explicitly modeling the "components" of definitions, leading to under-specific generation results. In this paper, we propose ESD, namely Explicit Semantic Decomposition for definition Generation, which explicitly decomposes the meaning of words into semantic components, and models them with discrete latent variables for definition generation. Experimental results show that ESD achieves top results on WordNet and Oxford benchmarks, outperforming strong previous baselines.

Improved Natural Language Generation via Loss Truncation [Website][PDF]
Daniel Kang and Tatsunori Hashimoto 6:00–7:00

Neural language models are usually trained to match the distributional properties of large-scale corpora by minimizing the log loss. While straightforward to optimize, this approach forces the model to reproduce all variations in the dataset, including noisy and invalid references (e.g., misannotations and hallucinated facts). Even a small fraction of noisy data can degrade the performance of log loss. As an alternative, prior work has shown that minimizing the distinguishability of generated samples is a principled and robust loss that can handle invalid references. However, distinguishability has not been used in practice due to challenges in optimization and estimation. We propose loss truncation: a simple and scalable procedure which adaptively removes high log loss examples as a way to optimize for distinguishability. Empirically, we demonstrate that loss truncation outperforms existing baselines on distinguishability on a summarization task. Furthermore, we show that samples generated by the loss truncation model have factual accuracy ratings that exceed those of baselines and match human references.

Line Graph Enhanced AMR-to-Text Generation with Mix-Order Graph Attention Networks [Website][PDF]
Yanbin Zhao, Lu Chen, Zhi Chen, Ruisheng Cao, Su Zhu, and Kai Yu 6:00–7:00

Efficient structure encoding for graphs with labeled edges is an important yet challenging point in many graph-based models. This work focuses on AMR-to-text generation – A graph-to-sequence task aiming to recover natural language from Abstract Meaning Representations (AMR). Existing graph-to-sequence approaches generally utilize graph neural networks as their encoders, which have two limitations: 1) The message propagation process in AMR graphs is only guided by the first-order adjacency information. 2) The relationships between labeled edges are not fully considered. In this work, we propose a novel graph encoding framework which can effectively explore the edge relations. We also adopt graph attention networks with higher-order neighborhood information to encode the rich structure in AMR graphs. Experiment results show that our approach obtains new state-of-the-art performance on English AMR benchmark datasets. The ablation analyses also demonstrate that both edge relations and higher-order information are beneficial to graph-to-sequence modeling.

Rigid Formats Controlled Text Generation [Website][PDF]
Piji Li, Haisong Zhang, Xiaojiang Liu, and Shuming Shi 6:00–7:00

Neural text generation has made tremendous progress in various tasks. One common characteristic of most of the tasks is that the texts are not restricted to some rigid formats when generating. However, we may confront some special text paradigms such as Lyrics (assume the music score is given), Sonnet, SongCi (classical Chinese poetry of the Song dynasty), etc. The typical characteristics of these texts are in three folds: (1) They must comply fully with the rigid predefined formats. (2) They must obey some rhyming schemes. (3) Although they are restricted to some formats, the sentence integrity must be guaranteed. To the best of our knowledge, text generation based on the predefined rigid formats has not been well investigated. Therefore, we propose a simple and elegant framework named SongNet to tackle this problem. The backbone of the framework is a Transformer-based auto-regressive language model. Sets of symbols are tailor-designed to improve the modeling performance especially on format, rhyme, and sentence integrity. We improve the attention mechanism to impel the model to capture some future information on the format. A pre-training and fine-tuning framework is designed to further improve the generation quality. Extensive experiments conducted on two collected corpora demonstrate that our proposed framework generates significantly better results in terms of both automatic metrics and the human evaluation.

Syn-QG: Syntactic and Shallow Semantic Rules for Question Generation [Website][PDF]
Kaustubh Dhole and Christopher D. Manning 6:00–7:00

Question Generation (QG) is fundamentally a simple syntactic transformation; however, many aspects of semantics influence what questions are good to form. We implement this observation by developing Syn-QG, a set of transparent syntactic rules leveraging universal dependencies, shallow semantic parsing, lexical resources, and custom rules which transform declarative sentences into question-answer pairs. We utilize PropBank argument descriptions and VerbNet state predicates to incorporate shallow semantic content, which helps generate questions of a descriptive nature and produce inferential and semantically richer questions than existing systems. In order to improve syntactic fluency and eliminate grammatically incorrect questions, we employ back-translation over the output of these syntactic rules. A set of crowd-sourced evaluations shows that our system can generate a larger number of highly grammatical and relevant questions than previous QG systems and that back-translation drastically improves grammaticality at a slight cost of generating irrelevant questions.

Session 1B: Information Retrieval and Text Mining-2

An Online Semantic-enhanced Dirichlet Model for Short Text Stream Clustering [Website][PDF]
Jay Kumar, Junming Shao, Salah Uddin, and Wazir Ali 6:00–7:00

Clustering short text streams is a challenging task due to its unique properties: infinite length, sparse data representation and cluster evolution. Existing approaches often exploit short text streams in a batch way. However, determine the optimal batch size is usually a difficult task since we have no priori knowledge when the topics evolve. In addition, traditional independent word representation in graphical model tends to cause "term ambiguity" problem in short text clustering. Therefore, in this paper, we propose an Online Semantic-enhanced Dirichlet Model for short sext stream clustering, called OSDM, which integrates the word-occurance semantic information (i.e., context) into a new graphical model and clusters each arriving short text automatically in an online way. Extensive results have demonstrated that OSDM has better performance compared to many state-of-the-art algorithms on both synthetic and real-world data sets.

Generative Semantic Hashing Enhanced via Boltzmann Machines [Website][PDF]
Lin Zheng, Qinliang Su, Dinghan Shen, and Changyou Chen 6:00–7:00

Generative semantic hashing is a promising technique for large-scale information retrieval thanks to its fast retrieval speed and small memory footprint. For the tractability of training, existing generative-hashing methods mostly assume a factorized form for the posterior distribution, enforcing independence among the bits of hash codes. From the perspectives of both model representation and code space size, independence is always not the best assumption. In this paper, to introduce correlations among the bits of hash codes, we propose to employ the distribution of Boltzmann machine as the variational posterior. To address the intractability issue of training, we first develop an approximate method to reparameterize the distribution of a Boltzmann machine by augmenting it as a hierarchical concatenation of a Gaussian-like distribution and a Bernoulli distribution. Based on that, an asymptotically-exact lower bound is further derived for the evidence lower bound (ELBO). With these novel techniques, the entire model can be optimized efficiently. Extensive experimental results demonstrate that by effectively modeling correlations among different bits within a hash code, our model can achieve significant performance gains.

Interactive Construction of User-Centric Dictionary for Text Analytics [Website][PDF]
Ryosuke Kohita, Issei Yoshida, Hiroshi Kanayama, and Tetsuya Nasukawa 6:00–7:00

We propose a methodology to construct a term dictionary for text analytics through an interactive process between a human and a machine, which helps the creation of flexible dictionaries with precise granularity required in typical text analysis. This paper introduces the first formulation of interactive dictionary construction to address this issue. To optimize the interaction, we propose a new algorithm that effectively captures an analyst's intention starting from only a small number of sample terms. Along with the algorithm, we also design an automatic evaluation framework that provides a systematic assessment of any interactive method for the dictionary creation task. Experiments using real scenario based corpora and dictionaries show that our algorithm outperforms baseline methods, and works even with a small number of interactions.

Tree-Structured Neural Topic Model [Website][PDF]
Masaru Isonuma, Junichiro Mori, Danushka Bollegala, and Ichiro Sakata 6:00–7:00

This paper presents a tree-structured neural topic model, which has a topic distribution over a tree with an infinite number of branches. Our model parameterizes an unbounded ancestral and fraternal topic distribution by applying doubly-recurrent neural networks. With the help of autoencoding variational Bayes, our model improves data scalability and achieves competitive performance when inducing latent topics and tree structures, as compared to a prior tree-structured topic model (Blei et al., 2010). This work extends the tree-structured topic model such that it can be incorporated with neural models for downstream tasks.

Unsupervised FAQ Retrieval with Question Generation and BERT [Website][PDF]
Yosi Mass, Boaz Carmeli, Haggai Roitman, and David Konopnicki 6:00–7:00

We focus on the task of Frequently Asked Questions (FAQ) retrieval. A given user query can be matched against the questions and/or the answers in the FAQ. We present a fully unsupervised method that exploits the FAQ pairs to train two BERT models. The two models match user queries to FAQ answers and questions, respectively. We alleviate the missing labeled data of the latter by automatically generating high-quality question paraphrases. We show that our model is on par and even outperforms supervised models on existing datasets.

Session 1B: NLP Applications-1

"The Boating Store Had Its Best Sail Ever": Pronunciation-attentive Contextualized Pun Recognition [Website][PDF]
Yichao Zhou, Jyun-Yu Jiang, Jieyu Zhao, Kai-Wei Chang, and Wei Wang 6:00–7:00

Humor plays an important role in human languages and it is essential to model humor when building intelligence systems. Among different forms of humor, puns perform wordplay for humorous effects by employing words with double entendre and high phonetic similarity. However, identifying and modeling puns are challenging as puns usually involved implicit semantic or phonological tricks. In this paper, we propose Pronunciation-attentive Contextualized Pun Recognition (PCPR) to perceive human humor, detect if a sentence contains puns and locate them in the sentence. PCPR derives contextualized representation for each word in a sentence by capturing the association between the surrounding context and its corresponding phonetic symbols. Extensive experiments are conducted on two benchmark datasets. Results demonstrate that the proposed approach significantly outperforms the state-of-the-art methods in pun detection and location tasks. In-depth analyses verify the effectiveness and robustness of PCPR.

Fast and Accurate Deep Bidirectional Language Representations for Unsupervised Learning [Website][PDF]
Joongbo Shin, Yoonhyung Lee, Seunghyun Yoon, and Kyomin Jung 6:00–7:00

Even though BERT has achieved successful performance improvements in various supervised learning tasks, BERT is still limited by repetitive inferences on unsupervised tasks for the computation of contextual language representations. To resolve this limitation, we propose a novel deep bidirectional language model called a Transformer-based Text Autoencoder (T-TA). The T-TA computes contextual language representations without repetition and displays the benefits of a deep bidirectional architecture, such as that of BERT. In computation time experiments in a CPU environment, the proposed T-TA performs over six times faster than the BERT-like model on a reranking task and twelve times faster on a semantic similarity task. Furthermore, the T-TA shows competitive or even better accuracies than those of BERT on the above tasks. Code is available at https://github.com/joongbo/tta.

Fine-grained Interest Matching for Neural News Recommendation [Website][PDF]
Heyuan Wang, Fangzhao Wu, Zheng Liu, and Xing Xie 6:00–7:00

Personalized news recommendation is a critical technology to improve users' online news reading experience. The core of news recommendation is accurate matching between user's interests and candidate news. The same user usually has diverse interests that are reflected in different news she has browsed. Meanwhile, important semantic features of news are implied in text segments of different granularities. Existing studies generally represent each user as a single vector and then match the candidate news vector, which may lose fine-grained information for recommendation. In this paper, we propose FIM, a Fine-grained Interest Matching method for neural news recommendation. Instead of aggregating user's all historical browsed news into a unified vector, we hierarchically construct multi-level representations for each news via stacked dilated convolutions. Then we perform fine-grained matching between segment pairs of each browsed news and the candidate news at each semantic level. High-order salient signals are then identified by resembling the hierarchy of image recognition for final click prediction. Extensive experiments on a real-world dataset from MSN news validate the effectiveness of our model on news recommendation.

Interpretable Operational Risk Classification with Semi-Supervised Variational Autoencoder [Website][PDF]
Fan Zhou, Shengming Zhang, and Yi Yang 6:00–7:00

Operational risk management is one of the biggest challenges nowadays faced by financial institutions. There are several major challenges of building a text classification system for automatic operational risk prediction, including imbalanced labeled/unlabeled data and lacking interpretability. To tackle these challenges, we present a semi-supervised text classification framework that integrates multi-head attention mechanism with Semi-supervised variational inference for Operational Risk Classification (SemiORC). We empirically evaluate the framework on a real-world dataset. The results demonstrate that our method can better utilize unlabeled data and learn visually interpretable document representations. SemiORC also outperforms other baseline methods on operational risk classification.

Interpreting Twitter User Geolocation [Website][PDF]
Ting Zhong, Tianliang Wang, Fan Zhou, Goce Trajcevski, Kunpeng Zhang, and Yi Yang 6:00–7:00

Identifying user geolocation in online social networks is an essential task in many location-based applications. Existing methods rely on the similarity of text and network structure, however, they suffer from a lack of interpretability on the corresponding results, which is crucial for understanding model behavior. In this work, we adopt influence functions to interpret the behavior of GNN-based models by identifying the importance of training users when predicting the locations of the testing users. This methodology helps with providing meaningful explanations on prediction results. Furthermore, it also initiates an attempt to uncover the so-called "black-box" GNN-based models by investigating the effect of individual nodes.

Modeling Code-Switch Languages Using Bilingual Parallel Corpus [Website][PDF]
Grandee Lee and Haizhou Li 6:00–7:00

Language modeling is the technique to estimate the probability of a sequence of words. A bilingual language model is expected to model the sequential dependency for words across languages, which is difficult due to the inherent lack of suitable training data as well as diverse syntactic structure across languages. We propose a bilingual attention language model (BALM) that simultaneously performs language modeling objective with a quasi-translation objective

to model both the monolingual as well as the cross-lingual sequential dependency. The attention mechanism learns the bilingual context from a parallel corpus. BALM achieves state-of-the-art performance on the SEAME code-switch database by reducing the perplexity of 20.5% over the best-reported result. We also apply BALM in bilingual lexicon induction, and language normalization tasks to validate the idea.

SpellGCN: Incorporating Phonological and Visual Similarities into Language Models for Chinese Spelling Check [Website] [PDF]
Xingyi Cheng, Weidi Xu, Kunlong Chen, Shaohua Jiang, Feng Wang, Taifeng Wang, Wei Chu, and Yuan Qi 6:00–7:00

Chinese Spelling Check (CSC) is a task to detect and correct spelling errors in Chinese natural language. Existing methods have made attempts to incorporate the similarity knowledge between Chinese characters. However, they take the similarity knowledge as either an external input resource or just heuristic rules. This paper proposes to incorporate phonological and visual similarity knowledge into language models for CSC via a specialized graph convolutional network (SpellGCN). The model builds a graph over the characters, and SpellGCN is learned to map this graph into a set of inter-dependent character classifiers. These classifiers are applied to the representations extracted by another network, such as BERT, enabling the whole network to be end-to-end trainable. Experiments are conducted on three human-annotated datasets. Our method achieves superior performance against previous models by a large margin.

Spelling Error Correction with Soft-Masked BERT [Website] [PDF]
Shaohua Zhang, Haoran Huang, Jicong Liu, and Hang Li 6:00–7:00

Spelling error correction is an important yet challenging task because a satisfactory solution of it essentially needs human-level language understanding ability. Without loss of generality we consider Chinese spelling error correction (CSC) in this paper. A state-of-the-art method for the task selects a character from a list of candidates for correction (including non-correction) at each position of the sentence on the basis of BERT, the language representation model. The accuracy of the method can be sub-optimal, however, because BERT does not have sufficient capability to detect whether there is an error at each position, apparently due to the way of pre-training it using mask language modeling. In this work, we propose a novel neural architecture to address the aforementioned issue, which consists of a network for error detection and a network for error correction based on BERT, with the former being connected to the latter with what we call soft-masking technique. Our method of using 'Soft-Masked BERT' is general, and it may be employed in other language detection-correction problems. Experimental results on two datasets, including one large dataset which we create and plan to release, demonstrate that the performance of our proposed method is significantly better than the baselines including the one solely based on BERT.

Session 1B: Question Answering-1

A Frame-based Sentence Representation for Machine Reading Comprehension [Website][PDF]
Shaoru Guo, Ru Li, Hongye Tan, Xiaoli Li, Yong Guan, Hongyan Zhao, and Yueping Zhang 6:00–7:00

Sentence representation (SR) is the most crucial and challenging task in Machine Reading Comprehension (MRC). MRC systems typically only utilize the information contained in the sentence itself, while human beings can leverage their semantic knowledge. To bridge the gap, we proposed a novel Frame-based Sentence Representation (FSR) method, which employs frame semantic knowledge to facilitate sentence modelling. Specifically, different from existing methods that only model lexical units (LUs), Frame Representation Models, which utilize both LUs in frame and Frame-to-Frame (F-to-F) relations, are designed to model frames and sentences with attention schema. Our proposed FSR method is able to integrate multiple-frame semantic information to get much better sentence representations. Our extensive experimental results show that it performs better than state-of-the-art technologies on machine reading comprehension task.

A Methodology for Creating Question Answering Corpora Using Inverse Data Annotation [Website][PDF]
Jan Deriu, Katsiaryna Mlynchyk, Philippe Schläpfer, Alvaro Rodrigo, Dirk von Grünigen, Nicolas Kaiser, Kurt Stockinger, Eneko Agirre, and Mark Cieliebak 6:00–7:00

In this paper, we introduce a novel methodology to efficiently construct a corpus for question answering over structured data. For this, we introduce an intermediate representation that is based on the logical query plan in a database, called Operation Trees (OT). This representation allows us to invert the annotation process without loosing flexibility in the types of queries that we generate. Furthermore, it allows for fine-grained alignment of the tokens to the operations. Thus, we randomly generate OTs from a context free grammar and annotators just have to write the appropriate question and assign the tokens. We compare our corpus OTTA (Operation Trees and Token Assignment), a large semantic parsing corpus for evaluating natural language interfaces to databases, to Spider and LC-QuaD 2.0 and show that our methodology more than triples the annotation speed while maintaining the complexity of the queries. Finally, we train a state-of-the-art semantic parsing model on our data and show that our dataset is a challenging dataset and that the token alignment can be leveraged to significantly increase the performance.

Contextualized Sparse Representations for Real-Time Open-Domain Question Answering [Website][PDF]
Jinhyuk Lee, Minjoon Seo, Hannaneh Hajishirzi, and Jaewoo Kang 6:00–7:00

Open-domain question answering can be formulated as a phrase retrieval problem, in which we can expect huge scalability and speed benefit but often suffer from low accuracy due to the limitation of existing phrase representation models. In this paper, we aim to improve the quality of each phrase embedding by augmenting it with a contextualized sparse representation (Sparc). Unlike previous sparse vectors that are term-frequency-based (e.g., tf-idf) or directly learned (only few thousand dimensions), we leverage rectified self-attention to indirectly learn sparse vectors in n-gram vocabulary space. By augmenting the previous phrase retrieval model (Seo et al., 2019) with Sparc, we show 4%+ improvement in CuratedTREC and SQuAD-Open. Our CuratedTREC score is even better than the best known retrieve & read model with at least 45x faster inference speed.

Dynamic Sampling Strategies for Multi-Task Reading Comprehension [Website][PDF]
Ananth Gottumukkala, Dheeru Dua, Sameer Singh, and Matt Gardner 6:00–7:00

Building general reading comprehension systems, capable of solving multiple datasets at the same time, is a recent aspirational goal in the research community. Prior work has focused on model architecture or generalization to held out datasets, and largely passed over the particulars of the multi-task learning set up. We show that a simple dynamic sampling strategy, selecting instances for training proportional to the multi-task model's current performance on a dataset relative to its single task performance, gives substantive gains over prior multi-task sampling strategies, mitigating the catastrophic forgetting that is common in multi-task learning. We also demonstrate that allowing instances of different tasks to be interleaved as much as possible between each epoch and batch has a clear benefit in multi-task performance over forcing task homogeneity at the epoch or batch level. Our final model shows greatly increased performance over the best model on ORB, a recently-released multitask reading comprehension benchmark.

Enhancing Answer Boundary Detection for Multilingual Machine Reading Comprehension [Website][PDF]
Fei Yuan, Linjun Shou, Xuanyu Bai, Ming Gong, Yaobo Liang, Nan Duan, Yan Fu, and Daxin Jiang 6:00–7:00

Multilingual pre-trained models could leverage the training data from a rich source language (such as English) to improve performance on low resource languages. However, the transfer quality for multilingual Machine Reading Comprehension (MRC) is significantly worse than sentence classification tasks mainly due to the requirement of MRC to detect the word level answer boundary. In this paper, we propose two auxiliary tasks in the fine-tuning stage to create additional phrase boundary supervision: (1) A mixed MRC task, which translates the question or passage to other languages and builds cross-lingual question-passage pairs; (2) A language-agnostic knowledge masking task by leveraging knowledge phrases mined from web. Besides, extensive experiments on two cross-lingual MRC datasets show the effectiveness of our proposed approach.

Explicit Memory Tracker with Coarse-to-Fine Reasoning for Conversational Machine Reading [Website][PDF]
Yifan Gao, Chien-Sheng Wu, Shafiq Joty, Caiming Xiong, Richard Socher, Irwin King, Michael Lyu, and Steven C.H. Hoi 6:00–7:00

The goal of conversational machine reading is to answer user questions given a knowledge base text which may require asking clarification questions. Existing approaches are limited in their decision making due to struggles in extracting question-related rules and reasoning about them. In this paper, we present a new framework of conversational machine reading that comprises a novel Explicit Memory Tracker (EMT) to track whether conditions listed in the rule text have already been satisfied to make a decision. Moreover, our framework generates clarification questions by adopting a coarse-to-fine reasoning strategy, utilizing sentence-level entailment scores to weight token-level distributions. On the ShARC benchmark (blind, held-out) testset, EMT achieves new state-of-the-art results of 74.6% micro-averaged decision accuracy and 49.5 BLEU4. We also show that EMT is more interpretable by visualizing the entailment-oriented reasoning process as the conversation flows. Code and models are released at https://github.com/Yifan-Gao/explicit_memory_tracker.

Injecting Numerical Reasoning Skills into Language Models [Website][PDF]
Mor Geva, Ankit Gupta, and Jonathan Berant 6:00–7:00

Large pre-trained language models (LMs) are known to encode substantial amounts of linguistic information. However, high-level reasoning skills, such as numerical reasoning, are difficult to learn from a language-modeling objective only. Consequently, existing models for numerical reasoning have used specialized architectures with limited flexibility. In this work, we show that numerical reasoning is amenable to automatic data generation, and thus one can inject this skill into pre-trained LMs, by generating large amounts of data, and training in a multi-task setup. We show that pre-training our model, GenBERT, on this data, dramatically improves performance on DROP (49.3 –> 72.3 F1), reaching performance that matches state-of-the-art models of comparable size, while using a simple and general-purpose encoder-decoder architecture. Moreover, GenBERT generalizes well to math word problem datasets, while maintaining high performance on standard RC tasks. Our approach provides a general recipe for injecting skills into large pre-trained LMs, whenever the skill is amenable to automatic data augmentation.

Learning to Identify Follow-Up Questions in Conversational Question Answering [Website][PDF]
Souvik Kundu, Qian Lin, and Hwee Tou Ng 6:00–7:00

Despite recent progress in conversational question answering, most prior work does not focus on follow-up questions. Practical conversational question answering systems often receive follow-up questions in an ongoing conversation, and it is crucial for a system to be able to determine whether a question is a follow-up question of the current conversation, for more effective answer finding subsequently. In this paper, we introduce a new follow-up question identification task. We propose a three-way attentive pooling network that determines the suitability of a follow-up question by capturing pair-wise interactions between the associated passage, the conversation history, and a candidate follow-up question. It enables the model to capture topic continuity and topic shift while scoring a particular candidate follow-up question. Experiments show that our proposed three-way attentive pooling network outperforms all baseline systems by significant margins.

Query Graph Generation for Answering Multi-hop Complex Questions from Knowledge Bases [Website][PDF]
Yunshi Lan and Jing Jiang 6:00–7:00

Previous work on answering complex questions from knowledge bases usually separately addresses two types of complexity: questions with constraints and questions with multiple hops of relations. In this paper, we handle both types of complexity at the same time. Motivated by the observation that early incorporation of constraints into query graphs can more effectively prune the search space, we propose a modified staged query graph generation method with more flexible ways to generate query graphs. Our experiments clearly show that our method achieves the state of the art on three benchmark KBQA datasets.

Session 1B: Resources and Evaluation-1

A Diverse Corpus for Evaluating and Developing English Math Word Problem Solvers [Website][PDF]
Shen-yun Miao, Chao-Chun Liang, and Keh-Yih Su 6:00–7:00

We present ASDiv (Academia Sinica Diverse MWP Dataset), a diverse (in terms of both language patterns and problem types) English math word problem (MWP) corpus for evaluating the capability of various MWP solvers. Existing MWP corpora for studying AI progress remain limited either in language usage patterns or in problem types. We thus present a new English MWP corpus with 2,305 MWPs that cover more text patterns and most problem types taught in elementary school. Each MWP is annotated with its problem type and grade level (for indicating the level of difficulty). Furthermore, we propose a metric to measure the lexicon usage diversity of a given MWP corpus, and demonstrate that ASDiv is more diverse than existing corpora. Experiments show that our proposed corpus reflects the true capability of MWP solvers more faithfully.

Improving Image Captioning Evaluation by Considering Inter References Variance [Website][PDF]
Yanzhi Yi, Hangyu Deng, and Jinglu Hu 6:00–7:00

Evaluating image captions is very challenging partially due to the fact that there are multiple correct captions for every single image. Most of the existing one-to-one metrics operate by penalizing mismatches between reference and generative caption without considering the intrinsic variance between ground truth captions. It usually leads to over-penalization and thus a bad correlation to human judgment. Recently, the latest one-to-one metric BERTScore can achieve high human correlation in system-level tasks while some issues can be fixed for better performance. In this paper, we propose a novel metric based on BERTScore that could handle such a challenge and extend BERTScore with a few new features appropriately for image captioning evaluation. The experimental results show that our metric achieves state-of-the-art human judgment correlation.

Revisiting the Context Window for Cross-lingual Word Embeddings [Website][PDF]
Ryokan Ri and Yoshimasa Tsuruoka 6:00–7:00

Existing approaches to mapping-based cross-lingual word embeddings are based on the assumption that the source and target embedding spaces are structurally similar. The structures of embedding spaces largely depend on the co-occurrence statistics of each word, which the choice of context window determines. Despite this obvious connection between the context window and mapping-based cross-lingual embeddings, their relationship has been underexplored in prior work. In this work, we provide a thorough evaluation, in various languages, domains, and tasks, of bilingual embeddings trained with different context windows. The highlight of our findings is that increasing the size of both the source and target window sizes improves the performance of bilingual lexicon induction, especially the performance on frequent nouns.

Session 1B Semantics: Lexical-1

Moving Down the Long Tail of Word Sense Disambiguation with Gloss Informed Bi-encoders [Website][PDF]

Terra Blevins and Luke Zettlemoyer 6:00–7:00

A major obstacle in Word Sense Disambiguation (WSD) is that word senses are not uniformly distributed, causing existing models to generally perform poorly on senses that are either rare or unseen during training. We propose a bi-encoder model that independently embeds (1) the target word with its surrounding context and (2) the dictionary definition, or gloss, of each sense. The encoders are jointly optimized in the same representation space, so that sense disambiguation can be performed by finding the nearest sense embedding for each target word embedding. Our system outperforms previous state-of-the-art models on English all-words WSD; these gains predominantly come from improved performance on rare senses, leading to a 31.1% error reduction on less frequent senses over prior work. This demonstrates that rare senses can be more effectively disambiguated by modeling their definitions.

Session 1B: Student Research Workshop

Grammatical Error Correction Using Pseudo Learner Corpus Considering Learner's Error Tendency
[Website] [PDF]
Yujin Takahashi, Satoru Katsumata, and Mamoru Komachi 6:00–7:00

Recently, several studies have focused on improving the performance of grammatical error correction (GEC) tasks using pseudo data. However, a large amount of pseudo data are required to train an accurate GEC model. To address the limitations of language and computational resources, we assume that introducing pseudo errors into sentences similar to those written by the language learners is more efficient, rather than incorporating random pseudo errors into monolingual data. In this regard, we study the effect of pseudo data on GEC task performance using two approaches. First, we extract sentences that are similar to the learners' sentences from monolingual data. Second, we generate realistic pseudo errors by considering error types that learners often make. Based on our comparative results, we observe that F0.5 scores for the Russian GEC task are significantly improved.

Research on Task Discovery for Transfer Learning in Deep Neural Networks [Website] [PDF]
Arda Akdemir 6:00–7:00

Deep neural network based machine learning models are shown to perform poorly on unseen or out-of-domain examples by numerous recent studies. Transfer learning aims to avoid overfitting and to improve generalizability by leveraging the information obtained from multiple tasks. Yet, the benefits of transfer learning depend largely on task selection and finding the right method of sharing. In this thesis, we hypothesize that current deep neural network based transfer learning models do not achieve their fullest potential for various tasks and there are still many task combinations that will benefit from transfer learning that are not considered by the current models. To this end, we started our research by implementing a novel multi-task learner with relaxed annotated data requirements and obtained a performance improvement on two NLP tasks. We will further devise models to tackle tasks from multiple areas of machine learning, such as Bioinformatics and Computer Vision, in addition to NLP.

RPD: A Distance Function Between Word Embeddings [Website] [PDF]
Xuhui Zhou, Shujian Huang, and Zaixiang Zheng 6:00–7:00

It is well-understood that different algorithms, training processes, and corpora produce different word embeddings. However, less is known about the relation between different embedding spaces, i.e. how far different sets of embeddings deviate from each other. In this paper, we propose a novel metric called Relative Pairwise Inner Product Distance (RPD) to quantify the distance between different sets of word embeddings. This unitary-invariant metric has a unified scale for comparing different sets of word embeddings. Based on the properties of RPD, we study the relations of word embeddings of different algorithms systematically and investigate the influence of different training processes and corpora. The results shed light on the poorly understood word embeddings and justify RPD as a measure of the distance of embedding space.

Reflection-based Word Attribute Transfer [Website] [PDF]
Yoichi Ishibashi, Katsuhito Sudoh, Koichiro Yoshino, and Satoshi Nakamura 6:00–7:00

Word embeddings, which often represent such analogic relations as king - man + woman ~ queen, can be used to change a word's attribute, including its gender. For transferring king into queen in this analogy-based manner, we subtract a difference vector man - woman based on the knowledge that king is male. However, developing such knowledge is very costly for words and attributes. In this work, we propose a novel method for word attribute transfer based on reflection mappings without such an analogy operation. Experimental results show that our proposed method can transfer the word attributes of the given words without changing the words that do not have the target attributes.

Demo Session 1C

Time: *6:30–7:15*

Syntactic Search by Example [Website][PDF]
Micah Shlain, Hillel Taub-Tabib, Shoval Sadde, and Yoav Goldberg

We present a system that allows a user to search a large linguistically annotated corpus using syntactic patterns over dependency graphs. In contrast to previous attempts to this effect, we introduce a light-weight query language that does not require the user to know the details of the underlying syntactic representations, and instead to query the corpus by providing an example sentence coupled with simple markup. Search is performed at an interactive speed due to efficient linguistic graph-indexing and retrieval engine. This allows for rapid exploration, development and refinement of syntax-based queries. We demonstrate the system using queries over two corpora: the English wikipedia, and a collection of English pubmed abstracts. A demo of the wikipedia system is available at https://allenai.github.io/spike/ .

Demo Session 2A

Time: *8:00–8:45*

Tabouid: a Wikipedia-based word guessing game [Website][PDF]
Timothée Bernard

We present Tabouid, a word-guessing game automatically generated from Wikipedia. Tabouid contains 10,000 (virtual) cards in English, and as many in French, covering not only words and linguistic expressions but also a variety of topics including artists, historical events or scientific concepts. Each card corresponds to a Wikipedia article, and conversely, any article could be turned into a card. A range of relatively simple NLP and machine-learning techniques are effectively integrated into a two-stage process. First, a large subset of Wikipedia articles are scored - this score estimates the difficulty, or alternatively, the playability of the page. Then, the best articles are turned into cards by selecting, for each of them, a list of banned words based on its content. We believe that the game we present is more than mere entertainment and that, furthermore, this paper has pedagogical potential.

Talk to Papers: Bringing Neural Question Answering to Academic Search [Website][PDF]
Tiancheng Zhao and Kyusong Lee

We introduce Talk to Papers, which exploits the recent open-domain question answering (QA) techniques to improve the current experience of academic search. It's designed to enable researchers to use natural language queries to find precise answers and extract insights from a massive amount of academic papers. We present a large improvement over classic search engine baseline on several standard QA datasets and provide the community a collaborative data collection tool to curate the first natural language processing research QA dataset via a community effort.

Session 2A Overview – Monday, July 6, 2020 8:00–9:00

Track A *Computational Social Science and Social Media-2* Abstracts	Code-Switching Patterns Can Be an Effective Route to Improve Performance of Downstream NLP Applications: A Case Study of Humour, Sarcasm and Hate Speech Detection *Bansal, Garimella, Suhane, Patro, and Mukherjee* [Website][PDF]	DTCA: Decision Tree-based Co-Attention Networks for Explainable Claim Verification *Wu, Rao, Liang, and Nazir* [Website][PDF]	Integrating Semantic and Structural Information with Graph Convolutional Network for Controversy Detection *Zhong, Cao, Sheng, Guo, and Wang* [Website][PDF]	Predicting the Topical Stance and Political Leaning of Media using Tweets *Stefanov, Darwish, Atanasov, and Nakov* [Website][PDF]	Simple, Interpretable and Stable Method for Detecting Words with Usage Change across Corpora *Gonen, Jawahar, Seddah, and Goldberg* [Website][PDF]
Track B *Dialogue and Interactive Systems-3* Abstracts	CDL: Curriculum Dual Learning for Emotion-Controllable Response Generation *Shen and Feng* [Website][PDF]	Learning Low-Resource End-To-End Goal-Oriented Dialog for Fast and Reliable System Deployment *Dai, Li, Tang, Li, Sun, and Zhu* [Website][PDF]	Response-Anticipated Memory for On-Demand Knowledge Integration in Response Generation *Tian, Bi, Lee, Xue, SONG, Liu, and Zhang* [Website][PDF]	Towards Conversational Recommendation over Multi-Type Dialogs *Liu, Wang, Niu, Wu, Che, and Liu* [Website][PDF]	Towards Unsupervised Language Understanding and Generation by Joint Dual Learning *Su, Huang, and Chen* [Website][PDF]
	Unknown Intent Detection Using Gaussian Mixture Model with an Application to Zero-shot Intent Classification *Yan, Fan, Li, Liu, Zhang, Wu, and Lam* [Website][PDF]				
Track C *Generation-3* Abstracts	Expertise Style Transfer: A New Task Towards Better Communication between Experts and Laymen *Cao, Shui, Pan, Kan, Liu, and Chua* [Website][PDF]	Fact-based Text Editing *Iso, Qiao, and Li* [Website][PDF]	Fluent Response Generation for Conversational Question Answering *Baheti, Ritter, and Small* [Website][PDF]	Learning to Ask More: Semi-Autoregressive Sequential Question Generation under Dual-Graph Interaction *Chai and Wan* [Website][PDF]	Line Graph Enhanced AMR-to-Text Generation with Mix-Order Graph Attention Networks *Zhao, Chen, Chen, Cao, Zhu, and Yu* [Website][PDF]
	Probabilistically Masked Language Model Capable of Autoregressive Generation in Arbitrary Word Order *Liao, Jiang, and Liu* [Website][PDF]	Review-based Question Generation with Adaptive Instance Transfer and Augmentation *Yu, Bing, Zhang, Lam, and Si* [Website][PDF]	Towards Faithful Neural Table-to-Text Generation with Content-Matching Constraints *Wang, Wang, An, Yu, and Chen* [Website][PDF]		

Track D *Information Retrieval and Text Mining-3* Abstracts	Dynamic Memory Induction Networks for Few-Shot Text Classification *Geng, Li, Li, Sun, and Zhu* [Website][PDF]	Every Document Owns Its Structure: Inductive Text Classification via Graph Neural Networks *Zhang, Yu, Cui, Wu, Wen, and Wang* [Website][PDF]	Exclusive Hierarchical Decoding for Deep Keyphrase Generation *Chen, Chan, Li, and King* [Website][PDF]	Hierarchy-Aware Global Model for Hierarchical Text Classification *Zhou, Ma, Long, Xu, Ding, Zhang, Xie, and Liu* [Website][PDF]	Interactive Construction of User-Centric Dictionary for Text Analytics *Kohita, Yoshida, Kanayama, and Nasukawa* [Website][PDF]
	Keyphrase Generation for Scientific Document Retrieval *Boudin, Gallina, and Aizawa* [Website][PDF]	Neural Topic Modeling with Bidirectional Adversarial Training *Wang, Hu, Zhou, He, Xiong, Ye, and Xu* [Website][PDF]	Text Classification with Negative Supervision *Ohashi, Takayama, Kajiwara, Chu, and Arase* [Website][PDF]	Tree-Structured Neural Topic Model *Isonuma, Mori, Bollegala, and Sakata* [Website][PDF]	
Track E *Phonology, Morphology and Word Segmentation-1* Abstracts	A Graph Autoencoder Model of Derivational Morphology *Hofmann, Schütze, and Pierrehumbert* [Website][PDF]				
Track F *Question Answering-2* Abstracts	A Frame-based Sentence Representation for Machine Reading Comprehension *Guo, Li, Tan, Li, Guan, Zhao, and Zhang* [Website][PDF]	A Methodology for Creating Question Answering Corpora Using Inverse Data Annotation *Deriu, Mlynchyk, Schläpfer, Rodrigo, Grünigen, Kaiser, Stockinger, Agirre, and Cieliebak* [Website][PDF]	Enhancing Answer Boundary Detection for Multilingual Machine Reading Comprehension *Yuan, Shou, Bai, Gong, Liang, Duan, Fu, and Jiang* [Website][PDF]	Explicit Memory Tracker with Coarse-to-Fine Reasoning for Conversational Machine Reading *Gao, Wu, Joty, Xiong, Socher, King, Lyu, and Hoi* [Website][PDF]	Injecting Numerical Reasoning Skills into Language Models *Geva, Gupta, and Berant* [Website][PDF]
	Learning to Identify Follow-Up Questions in Conversational Question Answering *Kundu, Lin, and Ng* [Website][PDF]	Query Graph Generation for Answering Multi-hop Complex Questions from Knowledge Bases *Lan and Jiang* [Website][PDF]			
Track G *Resources and Evaluation-2* Abstracts	Building a User-Generated Content North-African Arabizi Treebank: Tackling Hell *Seddah, Essaidi, Fethi, Futeral, Muller, Ortiz Suárez, Sagot, and Srivastava* [Website][PDF]	Crawling and Preprocessing Mailing Lists At Scale for Dialog Analysis *Bevendorff, Al Khatib, Potthast, and Stein* [Website][PDF]	Fine-Grained Analysis of Cross-Linguistic Syntactic Divergences *Nikolaev, Arviv, Karidi, Kenneth, Mitnik, Saeboe, and Abend* [Website][PDF]	Generating Counter Narratives against Online Hate Speech: Data and Strategies *Tekiroğlu, Chung, and Guerini* [Website][PDF]	KLEJ: Comprehensive Benchmark for Polish Language Understanding *Rybak, Mroczkowski, Tracz, and Gawlik* [Website][PDF]
	Learning and Evaluating Emotion Lexicons for 91 Languages *Buechel, Rücker, and Hahn* [Website][PDF]	Multi-Hypothesis Machine Translation Evaluation *Fomicheva, Specia, and Guzmán* [Website][PDF]	Multimodal Quality Estimation for Machine Translation *Okabe, Blain, and Specia* [Website][PDF]	PuzzLing Machines: A Challenge on Learning From Small Data *Şahin, Kementchedjhieva, Rust, and Gurevych* [Website][PDF]	The SOFC-Exp Corpus and Neural Approaches to Information Extraction in the Materials Science Domain *Friedrich, Adel, Tomazic, Hingerl, Benteau, Marusczyk, and Lange* [Website][PDF]

	The TechQA Dataset *Castelli, Chakravarti, Dana, Ferritto, Florian, Franz, Garg, Khandelwal, McCarley, McCawley, Nasr, Pan, Pendus, Pitrelli, Pujar, Roukos, Sakrajda, Sil, Uceda-Sosa, Ward, and Zhang* [Website] [PDF]	iSarcasm: A Dataset of Intended Sarcasm *Oprea and Magdy* [Website] [PDF]			
Track H *Sentence Level-1* Abstracts	AMR Parsing via Graph-Sequence Iterative Inference *Cai and Lam* [Website] [PDF]				
Track I *Student Research Workshop* Abstracts	Topic Balancing with Additive Regularization of Topic Models *Veselova and Vorontsov* [Website] [PDF]	Combining Subword Representations into Word-level Representations in the Transformer Architecture *Casas, Costa-jussà, and Fonollosa* [Website] [PDF]	Zero-shot North Korean to English Neural Machine Translation by Character Tokenization and Phoneme Decomposition *Kim, Hirasawa, and Komachi* [Website] [PDF]	Media Bias, the Social Sciences, and NLP: Automating Frame Analyses to Identify Bias by Word Choice and Labeling *Hamborg* [Website] [PDF]	
Track J *Summarization-1* Abstracts	A Large-Scale Multi-Document Summarization Dataset from the Wikipedia Current Events Portal *Gholipour Ghalandari, Hokamp, Pham, Glover, and Ifrim* [Website] [PDF]	Attend, Translate and Summarize: An Efficient Method for Neural Cross-Lingual Summarization *Zhu, Zhou, Zhang, and Zong* [Website] [PDF]	Examining the State-of-the-Art in News Timeline Summarization *Gholipour Ghalandari and Ifrim* [Website] [PDF]	Improving Truthfulness of Headline Generation *Matsumaru, Takase, and Okazaki* [Website] [PDF]	SUPERT: Towards New Frontiers in Unsupervised Evaluation Metrics for Multi-Document Summarization *Gao, Zhao, and Eger* [Website] [PDF]
	Self-Attention Guided Copy Mechanism for Abstractive Summarization *Xu, Li, Yuan, Wu, He, and Zhou* [Website] [PDF]				

Session 2A Details

Session 2A: Computational Social Science and Social Media-2

Code-Switching Patterns Can Be an Effective Route to Improve Performance of Downstream NLP Applications: A Case Study of Humour, Sarcasm and Hate Speech Detection [Website][PDF]
Srijan Bansal, Vishal Garimella, Ayush Suhane, Jasabanta Patro, and Animesh Mukherjee 8:00–9:00

In this paper, we demonstrate how code-switching patterns can be utilised to improve various downstream NLP applications. In particular, we encode various switching features to improve humour, sarcasm and hate speech detection tasks. We believe that this simple linguistic observation can also be potentially helpful in improving other similar NLP applications.

DTCA: Decision Tree-based Co-Attention Networks for Explainable Claim Verification [Website][PDF]
Lianwei Wu, Yuan Rao, yongqiang zhao yongqiang, Hao Liang, and Ambreen Nazir 8:00–9:00

Recently, many methods discover effective evidence from reliable sources by appropriate neural networks for explainable claim verification, which has been widely recognized. However, in these methods, the discovery process of evidence is nontransparent and unexplained. Simultaneously, the discovered evidence is aimed at the interpretability of the whole sequence of claims but insufficient to focus on the false parts of claims. In this paper, we propose a Decision Tree-based Co-Attention model (DTCA) to discover evidence for explainable claim verification. Specifically, we first construct Decision Tree-based Evidence model (DTE) to select comments with high credibility as evidence in a transparent and interpretable way. Then we design Co-attention Self-attention networks (CaSa) to make the selected evidence interact with claims, which is for 1) training DTE to determine the optimal decision thresholds and obtain more powerful evidence; and 2) utilizing the evidence to find the false parts in the claim. Experiments on two public datasets, RumourEval and PHEME, demonstrate that DTCA not only provides explanations for the results of claim verification but also achieves the state-of-the-art performance, boosting the F1-score by more than 3.11%, 2.41%, respectively.

Integrating Semantic and Structural Information with Graph Convolutional Network for Controversy Detection [Website][PDF]
Lei Zhong, Juan Cao, Qiang Sheng, Junbo Guo, and Ziang Wang 8:00–9:00

Identifying controversial posts on social media is a fundamental task for mining public sentiment, assessing the influence of events, and alleviating the polarized views. However, existing methods fail to 1) effectively incorporate the semantic information from content-related posts; 2) preserve the structural information for reply relationship modeling; 3) properly handle posts from topics dissimilar to those in the training set. To overcome the first two limitations, we propose Topic-Post-Comment Graph Convolutional Network (TPC-GCN), which integrates the information from the graph structure and content of topics, posts, and comments for post-level controversy detection. As to the third limitation, we extend our model to Disentangled TPC-GCN (DTPC-GCN), to disentangle topic-related and topic-unrelated features and then fuse dynamically. Extensive experiments on two real-world datasets demonstrate that our models outperform existing methods. Analysis of the results and cases proves that our models can integrate both semantic and structural information with significant generalizability.

Predicting the Topical Stance and Political Leaning of Media using Tweets [Website][PDF]
Peter Stefanov, Kareem Darwish, Atanas Atanasov, and Preslav Nakov 8:00–9:00

Discovering the stances of media outlets and influential people on current, debatable topics is important for social statisticians and policy makers. Many supervised solutions exist for determining viewpoints, but manually annotating training data is costly. In this paper, we propose a cascaded method that uses unsupervised learning to ascertain the stance of Twitter users with respect to a polarizing topic by leveraging their retweet behavior; then, it uses supervised learning based on user labels to characterize both the general political leaning of online media and of popular Twitter users, as well as their stance with respect to the target polarizing topic. We evaluate the model by comparing its predictions to gold labels from the Media Bias/Fact Check website, achieving 82.6% accuracy.

Simple, Interpretable and Stable Method for Detecting Words with Usage Change across Corpora [Website][PDF]
Hila Gonen, Ganesh Jawahar, Djamé Seddah, and Yoav Goldberg 8:00–9:00

The problem of comparing two bodies of text and searching for words that differ in their usage between them arises often in digital humanities and computational social science. This is commonly approached by training word embeddings on each corpus, aligning the vector spaces, and looking for words whose cosine distance in the aligned space is large. However, these methods often require extensive filtering of the vocabulary to perform well, and - as we show in this work - result in unstable, and hence less reliable, results. We propose an alternative approach that does not use vector space alignment, and instead considers the neighbors of each word. The method is simple, interpretable and stable. We demonstrate its effectiveness in 9 different setups, considering different corpus splitting criteria (age, gender and profession of tweet authors, time of tweet) and different languages (English, French and Hebrew).

Session 2A: Dialogue and Interactive Systems-3

CDL: Curriculum Dual Learning for Emotion-Controllable Response Generation [Website][PDF]
Lei Shen and Yang Feng 8:00–9:00

Emotion-controllable response generation is an attractive and valuable task that aims to make open-domain conversations more empathetic and engaging. Existing methods mainly enhance the emotion expression by adding regularization terms to standard cross-entropy loss and thus influence the training process. However, due to the lack of further consideration of content consistency, the common problem of response generation tasks, safe response, is intensified. Besides, query emotions that can help model the relationship between query and response are simply ignored in previous models, which would further hurt the coherence. To alleviate these problems, we propose a novel framework named Curriculum Dual Learning (CDL) which extends the emotion-controllable response generation to a dual task to generate emotional responses and emotional queries alternatively. CDL utilizes two rewards focusing on emotion and content to improve the duality. Additionally, it applies curriculum learning to gradually generate high-quality responses based on the difficulties of expressing various emotions. Experimental results show that CDL significantly outperforms the baselines in terms of coherence, diversity, and relation to emotion factors.

Learning Low-Resource End-To-End Goal-Oriented Dialog for Fast and Reliable System Deployment [Website][PDF]
Yinpei Dai, Hangyu Li, Chengguang Tang, Yongbin Li, Jian Sun, and Xiaodan Zhu 8:00–9:00

Existing end-to-end dialog systems perform less effectively when data is scarce. To obtain an acceptable success in real-life online services with only a handful of training examples, both fast adaptability and reliable performance are highly desirable for dialog systems. In this paper, we propose the Meta-Dialog System (MDS), which combines the advantages of both meta-learning approaches and human-machine collaboration. We evaluate our methods on a new extended-bAbI dataset and a transformed MultiWOZ dataset for low-resource goal-oriented dialog learning. Experimental results show that MDS significantly outperforms non-meta-learning baselines and can achieve more than 90% per-turn accuracies with only 10 dialogs on the extended-bAbI dataset.

Response-Anticipated Memory for On-Demand Knowledge Integration in Response Generation [Website][PDF]
Zhiliang Tian, Wei Bi, Dongkyu Lee, Lanqing Xue, YIPING SONG, Xiaojiang Liu, and Nevin L. Zhang 8:00–9:00

Neural conversation models are known to generate appropriate but non-informative responses in general. A scenario where informativeness can be significantly enhanced is Conversing by Reading (CbR), where conversations take place with respect to a given external document. In previous work, the external document is utilized by (1) creating a context-aware document memory that integrates information from the document and the conversational context, and then (2) generating responses referring to the memory. In this paper, we propose to create the document memory with some anticipated responses in mind. This is achieved using a teacher-student framework. The teacher is given the external document, the context, and the ground-truth response, and learns how to build a response-aware document memory from three sources of information. The student learns to construct a response-anticipated document memory from the first two sources, and teacher's insight on memory creation. Empirical results show that our model outperforms the previous state-of-the-art for the CbR task.

Towards Conversational Recommendation over Multi-Type Dialogs [Website][PDF]
Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu 8:00–9:00

We focus on the study of conversational recommendation in the context of multi-type dialogs, where the bots can proactively and naturally lead a conversation from a non-recommendation dialog (e.g., QA) to a recommendation dialog, taking into account user's interests and feedback. To facilitate the study of this task, we create a human-to-human Chinese dialog dataset DuRecDial (about 10k dialogs, 156k utterances), where there are multiple sequential dialogs for a pair of a recommendation seeker (user) and a recommender (bot). In each dialog, the recommender proactively leads a multi-type dialog to approach recommendation targets and then makes multiple recommendations with rich interaction behavior. This dataset allows us to systematically investigate different parts of the overall problem, e.g., how to naturally lead a dialog, how to interact with users for recommendation. Finally we establish baseline results on DuRecDial for future studies.

Towards Unsupervised Language Understanding and Generation by Joint Dual Learning [Website][PDF]
Shang-Yu Su, Chao-Wei Huang, and Yun-Nung Chen 8:00–9:00

In modular dialogue systems, natural language understanding (NLU) and natural language generation (NLG) are two critical components, where NLU extracts the semantics from the given texts and NLG is to construct corresponding natural language sentences based on the input semantic representations. However, the dual property between understanding and generation has been rarely explored. The prior work is the first attempt that utilized the duality between NLU and NLG to improve the performance via a dual supervised learning framework. However, the prior work still learned both components in a supervised manner; instead, this paper introduces a general learning framework to effectively exploit such duality, providing flexibility of incorporating both supervised and unsupervised learning algorithms to train language understanding and generation models in a joint fashion. The benchmark experiments demonstrate that the proposed approach is capable of boosting the performance of both NLU and NLG. The source code is available at: https://github.com/MiuLab/DuaLUG.

Unknown Intent Detection Using Gaussian Mixture Model with an Application to Zero-shot Intent Classification [Website] [PDF]

Guangfeng Yan, Lu Fan, Qimai Li, Han Liu, Xiaotong Zhang, Xiao-Ming Wu, and Albert Y.S. Lam 8:00–9:00

User intent classification plays a vital role in dialogue systems. Since user intent may frequently change over time in many realistic scenarios, unknown (new) intent detection has become an essential problem, where the study has just begun. This paper proposes a semantic-enhanced Gaussian mixture model (SEG) for unknown intent detection. In particular, we model utterance embeddings with a Gaussian mixture distribution and inject dynamic class semantic information into Gaussian means, which enables learning more class-concentrated embeddings that help to facilitate downstream outlier detection. Coupled with a density-based outlier detection algorithm, SEG achieves competitive results on three real task-oriented dialogue datasets in two languages for unknown intent detection. On top of that, we propose to integrate SEG as an unknown intent identifier into existing generalized zero-shot intent classification models to improve their performance. A case study on a state-of-the-art method, ReCapsNet, shows that SEG can push the classification performance to a significantly higher level.

Session 2A: Generation-3

Expertise Style Transfer: A New Task Towards Better Communication between Experts and Laymen [Website][PDF]
Yixin Cao, Ruihao Shui, Liangming Pan, Min-Yen Kan, Zhiyuan Liu, and Tat-Seng Chua 8:00–9:00

The curse of knowledge can impede communication between experts and laymen. We propose a new task of expertise style transfer and contribute a manually annotated dataset with the goal of alleviating such cognitive biases. Solving this task not only simplifies the professional language, but also improves the accuracy and expertise level of laymen descriptions using simple words. This is a challenging task, unaddressed in previous work, as it requires the models to have expert intelligence in order to modify text with a deep understanding of domain knowledge and structures. We establish the benchmark performance of five state-of-the-art models for style transfer and text simplification. The results demonstrate a significant gap between machine and human performance. We also discuss the challenges of automatic evaluation, to provide insights into future research directions. The dataset is publicly available at https://srhthu.github.io/expertise-style-transfer/.

Fact-based Text Editing [Website][PDF]
Hayate Iso, Chao Qiao, and Hang Li 8:00–9:00

We propose a novel text editing task, referred to as *fact-based text editing*, in which the goal is to revise a given document to better describe the facts in a knowledge base (e.g., several triples). The task is important in practice because reflecting the truth is a common requirement in text editing. First, we propose a method for automatically generating a dataset for research on fact-based text editing, where each instance consists of a draft text, a revised text, and several facts represented in triples. We apply the method into two public table-to-text datasets, obtaining two new datasets consisting of 233k and 37k instances, respectively. Next, we propose a new neural network architecture for fact-based text editing, called FACTEDITOR, which edits a draft text by referring to given facts using a buffer, a stream, and a memory. A straightforward approach to address the problem would be to employ an encoder-decoder model. Our experimental results on the two datasets show that FACTEDITOR outperforms the encoder-decoder approach in terms of fidelity and fluency. The results also show that FACTEDITOR conducts inference faster than the encoder-decoder approach.

Fluent Response Generation for Conversational Question Answering [Website][PDF]
Ashutosh Baheti, Alan Ritter, and Kevin Small 8:00–9:00

Question answering (QA) is an important aspect of open-domain conversational agents, garnering specific research focus in the conversational QA (ConvQA) subtask. One notable limitation of recent ConvQA efforts is the response being answer span extraction from the target corpus, thus ignoring the natural language generation (NLG) aspect of high-quality conversational agents. In this work, we propose a method for situating QA responses within a SEQ2SEQ NLG approach to generate fluent grammatical answer responses while maintaining correctness. From a technical perspective, we use data augmentation to generate training data for an end-to-end system. Specifically, we develop Syntactic Transformations (STs) to produce question-specific candidate answer responses and rank them using a BERT-based classifier (Devlin et al., 2019). Human evaluation on SQuAD 2.0 data (Rajpurkar et al., 2018) demonstrate that the proposed model outperforms baseline CoQA and QuAC models in generating conversational responses. We further show our model's scalability by conducting tests on the CoQA dataset. The code and data are available at https://github.com/abaheti95/QADialogSystem.

Learning to Ask More: Semi-Autoregressive Sequential Question Generation under Dual-Graph Interaction [Website][PDF]
Zi Chai and Xiaojun Wan 8:00–9:00

Traditional Question Generation (TQG) aims to generate a question given an input passage and an answer. When there is a sequence of answers, we can perform Sequential Question Generation (SQG) to produce a series of interconnected questions. Since the frequently occurred information omission and coreference between questions, SQG is rather challenging. Prior works regarded SQG as a dialog generation task and recurrently produced each question. However, they suffered from problems caused by error cascades and could only capture limited context dependencies. To this end, we generate questions in a semi-autoregressive way. Our model divides questions into different groups and generates each group of them in parallel. During this process, it builds two graphs focusing on information from passages, answers respectively and performs dual-graph interaction to get information for generation. Besides, we design an answer-aware attention mechanism and the coarse-to-fine generation scenario. Experiments on our new dataset containing 81.9K questions show that our model substantially outperforms prior works.

Line Graph Enhanced AMR-to-Text Generation with Mix-Order Graph Attention Networks [Website][PDF]
Yanbin Zhao, Lu Chen, Zhi Chen, Ruisheng Cao, Su Zhu, and Kai Yu 8:00–9:00

Efficient structure encoding for graphs with labeled edges is an important yet challenging point in many graph-based models. This work focuses on AMR-to-text generation – A graph-to-sequence task aiming to recover natural language from Abstract Meaning Representations (AMR). Existing graph-to-sequence approaches generally utilize graph neural networks as their encoders, which have two limitations: 1) The message propagation process in AMR graphs is only guided by the first-order adjacency information. 2) The relationships between labeled edges are not fully considered. In this work, we propose a novel graph encoding framework which can effectively explore the edge relations. We also adopt graph attention networks with higher-order neighborhood information to encode the rich structure in AMR graphs. Experiment results show that our approach obtains new state-of-the-art performance on English AMR

benchmark datasets. The ablation analyses also demonstrate that both edge relations and higher-order information are beneficial to graph-to-sequence modeling.

Probabilistically Masked Language Model Capable of Autoregressive Generation in Arbitrary Word Order [Website][PDF]
Yi Liao, Xin Jiang, and Qun Liu 8:00–9:00

Masked language model and autoregressive language model are two types of language models. While pretrained masked language models such as BERT overwhelm the line of natural language understanding (NLU) tasks, autoregressive language models such as GPT are especially capable in natural language generation (NLG). In this paper, we propose a probabilistic masking scheme for the masked language model, which we call probabilistically masked language model (PMLM). We implement a specific PMLM with a uniform prior distribution on the masking ratio named u-PMLM. We prove that u-PMLM is equivalent to an autoregressive permutated language model. One main advantage of the model is that it supports text generation in arbitrary order with surprisingly good quality, which could potentially enable new applications over traditional unidirectional generation. Besides, the pretrained u-PMLM also outperforms BERT on a bunch of downstream NLU tasks.

Review-based Question Generation with Adaptive Instance Transfer and Augmentation [Website][PDF]
Qian Yu, Lidong Bing, Qiong Zhang, Wai Lam, and Luo Si 8:00–9:00

While online reviews of products and services become an important information source, it remains inefficient for potential consumers to exploit verbose reviews for fulfilling their information need. We propose to explore question generation as a new way of review information exploitation, namely generating questions that can be answered by the corresponding review sentences. One major challenge of this generation task is the lack of training data, i.e. explicit mapping relation between the user-posed questions and review sentences. To obtain proper training instances for the generation model, we propose an iterative learning framework with adaptive instance transfer and augmentation. To generate to the point questions about the major aspects in reviews, related features extracted in an unsupervised manner are incorporated without the burden of aspect annotation. Experiments on data from various categories of a popular E-commerce site demonstrate the effectiveness of the framework, as well as the potentials of the proposed review-based question generation task.

Towards Faithful Neural Table-to-Text Generation with Content-Matching Constraints [Website][PDF]
Zhenyi Wang, Xiaoyang Wang, Bang An, Dong Yu, and Changyou Chen 8:00–9:00

Text generation from a knowledge base aims to translate knowledge triples to natural language descriptions. Most existing methods ignore the faithfulness between a generated text description and the original table, leading to generated information that goes beyond the content of the table. In this paper, for the first time, we propose a novel Transformer-based generation framework to achieve the goal. The core techniques in our method to enforce faithfulness include a new table-text optimal-transport matching loss and a table-text embedding similarity loss based on the Transformer model. Furthermore, to evaluate faithfulness, we propose a new automatic metric specialized to the table-to-text generation problem. We also provide detailed analysis on each component of our model in our experiments. Automatic and human evaluations show that our framework can significantly outperform state-of-the-art by a large margin.

Session 2A: Information Retrieval and Text Mining-3

Dynamic Memory Induction Networks for Few-Shot Text Classification [Website][PDF]
Ruiying Geng, Binhua Li, Yongbin Li, Jian Sun, and Xiaodan Zhu 8:00–9:00

This paper proposes Dynamic Memory Induction Networks (DMIN) for few-short text classification. The model develops a dynamic routing mechanism over static memory, enabling it to better adapt to unseen classes, a critical capability for few-short classification. The model also expands the induction process with supervised learning weights and query information to enhance the generalization ability of meta-learning. The proposed model brings forward the state-of-the-art performance significantly by 2~4% improvement on the miniRCV1 and ODIC datasets. Detailed analysis is further performed to show how the proposed network achieves the new performance.

Every Document Owns Its Structure: Inductive Text Classification via Graph Neural Networks [Website][PDF]
Yufeng Zhang, Xueli Yu, Zeyu Cui, Shu Wu, Zhongzhen Wen, and Liang Wang 8:00–9:00

Text classification is fundamental in natural language processing (NLP) and Graph Neural Networks (GNN) are recently applied in this task. However, the existing graph-based works can neither capture the contextual word relationships within each document nor fulfil the inductive learning of new words. Therefore in this work, to overcome such problems, we propose TextING for inductive text classification via GNN. We first build individual graphs for each document and then use GNN to learn the fine-grained word representations based on their local structure, which can also effectively produce embeddings for unseen words in the new document. Finally, the word nodes are aggregated as the document embedding. Extensive experiments on four benchmark datasets show that our method outperforms state-of-the-art text classification methods.

Exclusive Hierarchical Decoding for Deep Keyphrase Generation [Website][PDF]
Wang Chen, Hou Pong Chan, Piji Li, and Irwin King 8:00–9:00

Keyphrase generation (KG) aims to summarize the main ideas of a document into a set of keyphrases. A new setting is recently introduced into this problem, in which, given a document, the model needs to predict a set of keyphrases and simultaneously determine the appropriate number of keyphrases to produce. Previous work in this setting employs a sequential decoding process to generate keyphrases. However, such a decoding method ignores the intrinsic hierarchical compositionality existing in the keyphrase set of a document. Moreover, previous work tends to generate duplicated keyphrases, which wastes time and computing resources. To overcome these limitations, we propose an exclusive hierarchical decoding framework that includes a hierarchical decoding process and either a soft or a hard exclusion mechanism. The hierarchical decoding process is to explicitly model the hierarchical compositionality of a keyphrase set. Both the soft and the hard exclusion mechanisms keep track of previously-predicted keyphrases within a window size to enhance the diversity of the generated keyphrases. Extensive experiments on multiple KG benchmark datasets demonstrate the effectiveness of our method to generate less duplicated and more accurate keyphrases.

Hierarchy-Aware Global Model for Hierarchical Text Classification [Website][PDF]
Jie Zhou, Chunping Ma, Dingkun Long, Guangwei Xu, Ning Ding, Haoyu Zhang, Pengjun Xie, and Gongshen Liu 8:00–9:00

Hierarchical text classification is an essential yet challenging subtask of multi-label text classification with a taxonomic hierarchy. Existing methods have difficulties in modeling the hierarchical label structure in a global view. Furthermore, they cannot make full use of the mutual interactions between the text feature space and the label space. In this paper, we formulate the hierarchy as a directed graph and introduce hierarchy-aware structure encoders for modeling label dependencies. Based on the hierarchy encoder, we propose a novel end-to-end hierarchy-aware global model (HiAGM) with two variants. A multi-label attention variant (HiAGM-LA) learns hierarchy-aware label embeddings through the hierarchy encoder and conducts inductive fusion of label-aware text features. A text feature propagation model (HiAGM-TP) is proposed as the deductive variant that directly feeds text features into hierarchy encoders. Compared with previous works, both HiAGM-LA and HiAGM-TP achieve significant and consistent improvements on three benchmark datasets.

Interactive Construction of User-Centric Dictionary for Text Analytics [Website][PDF]
Ryosuke Kohita, Issei Yoshida, Hiroshi Kanayama, and Tetsuya Nasukawa 8:00–9:00

We propose a methodology to construct a term dictionary for text analytics through an interactive process between a human and a machine, which helps the creation of flexible dictionaries with precise granularity required in typical text analysis. This paper introduces the first formulation of interactive dictionary construction to address this issue. To optimize the interaction, we propose a new algorithm that effectively captures an analyst's intention starting from only a small number of sample terms. Along with the algorithm, we also design an automatic evaluation framework that provides a systematic assessment of any interactive method for the dictionary creation task. Experiments using real scenario based corpora and dictionaries show that our algorithm outperforms baseline methods, and works even with a small number of interactions.

Keyphrase Generation for Scientific Document Retrieval [Website][PDF]
Florian Boudin, Ygor Gallina, and Akiko Aizawa 8:00–9:00

Sequence-to-sequence models have lead to significant progress in keyphrase generation, but it remains unknown whether they are reliable enough to be beneficial for document retrieval. This study provides empirical evidence that such models can significantly improve retrieval performance, and introduces a new extrinsic evaluation framework that allows for a better understanding of the limitations of keyphrase generation models. Using this frame-

work, we point out and discuss the difficulties encountered with supplementing documents with -not present in text- keyphrases, and generalizing models across domains. Our code is available at https://github.com/boudinfl/ir-using-kg

Neural Topic Modeling with Bidirectional Adversarial Training [Website][PDF]
Rui Wang, Xuemeng Hu, Deyu Zhou, Yulan He, Yuxuan Xiong, Chenchen Ye, and Haiyang Xu 8:00–9:00

Recent years have witnessed a surge of interests of using neural topic models for automatic topic extraction from text, since they avoid the complicated mathematical derivations for model inference as in traditional topic models such as Latent Dirichlet Allocation (LDA). However, these models either typically assume improper prior (e.g. Gaussian or Logistic Normal) over latent topic space or could not infer topic distribution for a given document. To address these limitations, we propose a neural topic modeling approach, called Bidirectional Adversarial Topic (BAT) model, which represents the first attempt of applying bidirectional adversarial training for neural topic modeling. The proposed BAT builds a two-way projection between the document-topic distribution and the document-word distribution. It uses a generator to capture the semantic patterns from texts and an encoder for topic inference. Furthermore, to incorporate word relatedness information, the Bidirectional Adversarial Topic model with Gaussian (Gaussian-BAT) is extended from BAT. To verify the effectiveness of BAT and Gaussian-BAT, three benchmark corpora are used in our experiments. The experimental results show that BAT and Gaussian-BAT obtain more coherent topics, outperforming several competitive baselines. Moreover, when performing text clustering based on the extracted topics, our models outperform all the baselines, with more significant improvements achieved by Gaussian-BAT where an increase of near 6% is observed in accuracy.

Text Classification with Negative Supervision [Website][PDF]
Sora Ohashi, Junya Takayama, Tomoyuki Kajiwara, Chenhui Chu, and Yuki Arase 8:00–9:00

Advanced pre-trained models for text representation have achieved state-of-the-art performance on various text classification tasks. However, the discrepancy between the semantic similarity of texts and labelling standards affects classifiers, i.e. leading to lower performance in cases where classifiers should assign different labels to semantically similar texts. To address this problem, we propose a simple multitask learning model that uses negative supervision. Specifically, our model encourages texts with different labels to have distinct representations. Comprehensive experiments show that our model outperforms the state-of-the-art pre-trained model on both single- and multi-label classifications, sentence and document classifications, and classifications in three different languages.

Tree-Structured Neural Topic Model [Website][PDF]
Masaru Isonuma, Junichiro Mori, Danushka Bollegala, and Ichiro Sakata 8:00–9:00

This paper presents a tree-structured neural topic model, which has a topic distribution over a tree with an infinite number of branches. Our model parameterizes an unbounded ancestral and fraternal topic distribution by applying doubly-recurrent neural networks. With the help of autoencoding variational Bayes, our model improves data scalability and achieves competitive performance when inducing latent topics and tree structures, as compared to a prior tree-structured topic model (Blei et al., 2010). This work extends the tree-structured topic model such that it can be incorporated with neural models for downstream tasks.

Session 2A: Phonology, Morphology and Word Segmentation-1

A Graph Auto-encoder Model of Derivational Morphology [Website] [PDF]
Valentin Hofmann, Hinrich Schütze, and Janet Pierrehumbert 8:00–9:00

There has been little work on modeling the morphological well-formedness (MWF) of derivatives, a problem judged to be complex and difficult in linguistics. We present a graph auto-encoder that learns embeddings capturing information about the compatibility of affixes and stems in derivation. The auto-encoder models MWF in English surprisingly well by combining syntactic and semantic information with associative information from the mental lexicon.

Session 2A: Question Answering-2

A Frame-based Sentence Representation for Machine Reading Comprehension [Website][PDF]
Shaoru Guo, Ru Li, Hongye Tan, Xiaoli Li, Yong Guan, Hongyan Zhao, and Yueping Zhang 8:00–9:00

Sentence representation (SR) is the most crucial and challenging task in Machine Reading Comprehension (MRC). MRC systems typically only utilize the information contained in the sentence itself, while human beings can leverage their semantic knowledge. To bridge the gap, we proposed a novel Frame-based Sentence Representation (FSR) method, which employs frame semantic knowledge to facilitate sentence modelling. Specifically, different from existing methods that only model lexical units (LUs), Frame Representation Models, which utilize both LUs in frame and Frame-to-Frame (F-to-F) relations, are designed to model frames and sentences with attention schema. Our proposed FSR method is able to integrate multiple-frame semantic information to get much better sentence representations. Our extensive experimental results show that it performs better than state-of-the-art technologies on machine reading comprehension task.

A Methodology for Creating Question Answering Corpora Using Inverse Data Annotation [Website][PDF]
Jan Deriu, Katsiaryna Mlynchyk, Philippe Schläpfer, Alvaro Rodrigo, Dirk von Grünigen, Nicolas Kaiser, Kurt Stockinger, Eneko Agirre, and Mark Cieliebak 8:00–9:00

In this paper, we introduce a novel methodology to efficiently construct a corpus for question answering over structured data. For this, we introduce an intermediate representation that is based on the logical query plan in a database, called Operation Trees (OT). This representation allows us to invert the annotation process without loosing flexibility in the types of queries that we generate. Furthermore, it allows for fine-grained alignment of the tokens to the operations. Thus, we randomly generate OTs from a context free grammar and annotators just have to write the appropriate question and assign the tokens. We compare our corpus OTTA (Operation Trees and Token Assignment), a large semantic parsing corpus for evaluating natural language interfaces to databases, to Spider and LC-QuaD 2.0 and show that our methodology more than triples the annotation speed while maintaining the complexity of the queries. Finally, we train a state-of-the-art semantic parsing model on our data and show that our dataset is a challenging dataset and that the token alignment can be leveraged to significantly increase the performance.

Enhancing Answer Boundary Detection for Multilingual Machine Reading Comprehension [Website][PDF]
Fei Yuan, Linjun Shou, Xuanyu Bai, Ming Gong, Yaobo Liang, Nan Duan, Yan Fu, and Daxin Jiang 8:00–9:00

Multilingual pre-trained models could leverage the training data from a rich source language (such as English) to improve performance on low resource languages. However, the transfer quality for multilingual Machine Reading Comprehension (MRC) is significantly worse than sentence classification tasks mainly due to the requirement of MRC to detect the word level answer boundary. In this paper, we propose two auxiliary tasks in the fine-tuning stage to create additional phrase boundary supervision: (1) A mixed MRC task, which translates the question or passage to other languages and builds cross-lingual question-passage pairs; (2) A language-agnostic knowledge masking task by leveraging knowledge phrases mined from web. Besides, extensive experiments on two cross-lingual MRC datasets show the effectiveness of our proposed approach.

Explicit Memory Tracker with Coarse-to-Fine Reasoning for Conversational Machine Reading [Website][PDF]
Yifan Gao, Chien-Sheng Wu, Shafiq Joty, Caiming Xiong, Richard Socher, Irwin King, Michael Lyu, and Steven C.H. Hoi 8:00–9:00

The goal of conversational machine reading is to answer user questions given a knowledge base text which may require asking clarification questions. Existing approaches are limited in their decision making due to struggles in extracting question-related rules and reasoning about them. In this paper, we present a new framework of conversational machine reading that comprises a novel Explicit Memory Tracker (EMT) to track whether conditions listed in the rule text have already been satisfied to make a decision. Moreover, our framework generates clarification questions by adopting a coarse-to-fine reasoning strategy, utilizing sentence-level entailment scores to weight token-level distributions. On the ShARC benchmark (blind, held-out) testset, EMT achieves new state-of-the-art results of 74.6% micro-averaged decision accuracy and 49.5 BLEU4. We also show that EMT is more interpretable by visualizing the entailment-oriented reasoning process as the conversation flows. Code and models are released at https://github.com/Yifan-Gao/explicit_memory_tracker.

Injecting Numerical Reasoning Skills into Language Models [Website][PDF]
Mor Geva, Ankit Gupta, and Jonathan Berant 8:00–9:00

Large pre-trained language models (LMs) are known to encode substantial amounts of linguistic information. However, high-level reasoning skills, such as numerical reasoning, are difficult to learn from a language-modeling objective only. Consequently, existing models for numerical reasoning have used specialized architectures with limited flexibility. In this work, we show that numerical reasoning is amenable to automatic data generation, and thus one can inject this skill into pre-trained LMs, by generating large amounts of data, and training in a multi-task setup. We show that pre-training our model, GenBERT, on this data, dramatically improves performance on DROP (49.3 –> 72.3 F1), reaching performance that matches state-of-the-art models of comparable size, while using a simple and general-purpose encoder-decoder architecture. Moreover, GenBERT generalizes well to math word problem datasets, while maintaining high performance on standard RC tasks. Our approach provides a general recipe for injecting skills

into large pre-trained LMs, whenever the skill is amenable to automatic data augmentation.

Learning to Identify Follow-Up Questions in Conversational Question Answering [Website][PDF]
Souvik Kundu, Qian Lin, and Hwee Tou Ng 8:00–9:00

Despite recent progress in conversational question answering, most prior work does not focus on follow-up questions. Practical conversational question answering systems often receive follow-up questions in an ongoing conversation, and it is crucial for a system to be able to determine whether a question is a follow-up question of the current conversation, for more effective answer finding subsequently. In this paper, we introduce a new follow-up question identification task. We propose a three-way attentive pooling network that determines the suitability of a follow-up question by capturing pair-wise interactions between the associated passage, the conversation history, and a candidate follow-up question. It enables the model to capture topic continuity and topic shift while scoring a particular candidate follow-up question. Experiments show that our proposed three-way attentive pooling network outperforms all baseline systems by significant margins.

Query Graph Generation for Answering Multi-hop Complex Questions from Knowledge Bases [Website][PDF]
Yunshi Lan and Jing Jiang 8:00–9:00

Previous work on answering complex questions from knowledge bases usually separately addresses two types of complexity: questions with constraints and questions with multiple hops of relations. In this paper, we handle both types of complexity at the same time. Motivated by the observation that early incorporation of constraints into query graphs can more effectively prune the search space, we propose a modified staged query graph generation method with more flexible ways to generate query graphs. Our experiments clearly show that our method achieves the state of the art on three benchmark KBQA datasets.

Session 2A: Resources and Evaluation-2

Building a User-Generated Content North-African Arabizi Treebank: Tackling Hell [Website] [PDF]
Djamé Seddah, Farah Essaidi, Amal Fethi, Matthieu Futeral, Benjamin Muller, Pedro Javier Ortiz Suárez, Benoît Sagot, and Abhishek Srivastava 8:00–9:00

We introduce the first treebank for a romanized user-generated content variety of Algerian, a North-African Arabic dialect known for its frequent usage of code-switching. Made of 1500 sentences, fully annotated in morpho-syntax and Universal Dependency syntax, with full translation at both the word and the sentence levels, this treebank is made freely available. It is supplemented with 50k unlabeled sentences collected from Common Crawl and web-crawled data using intensive data-mining techniques. Preliminary experiments demonstrate its usefulness for POS tagging and dependency parsing. We believe that what we present in this paper is useful beyond the low-resource language community. This is the first time that enough unlabeled and annotated data is provided for an emerging user-generated content dialectal language with rich morphology and code switching, making it an challenging test-bed for most recent NLP approaches.

Crawling and Preprocessing Mailing Lists At Scale for Dialog Analysis [Website] [PDF]
Janek Bevendorff, Khalid Al Khatib, Martin Potthast, and Benno Stein 8:00–9:00

This paper introduces the Webis Gmane Email Corpus 2019, the largest publicly available and fully preprocessed email corpus to date. We crawled more than 153 million emails from 14,699 mailing lists and segmented them into semantically consistent components using a new neural segmentation model. With 96% accuracy on 15 classes of email segments, our model achieves state-of-the-art performance while being more efficient to train than previous ones. All data, code, and trained models are made freely available alongside the paper.

Fine-Grained Analysis of Cross-Linguistic Syntactic Divergences [Website] [PDF]
Dmitry Nikolaev, Ofir Arviv, Taelin Karidi, Neta Kenneth, Veronika Mitnik, Lilja Maria Saeboe, and Omri Abend 8:00–9:00

The patterns in which the syntax of different languages converges and diverges are often used to inform work on cross-lingual transfer. Nevertheless, little empirical work has been done on quantifying the prevalence of different syntactic divergences across language pairs. We propose a framework for extracting divergence patterns for any language pair from a parallel corpus, building on Universal Dependencies. We show that our framework provides a detailed picture of cross-language divergences, generalizes previous approaches, and lends itself to full automation. We further present a novel dataset, a manually word-aligned subset of the Parallel UD corpus in five languages, and use it to perform a detailed corpus study. We demonstrate the usefulness of the resulting analysis by showing that it can help account for performance patterns of a cross-lingual parser.

Generating Counter Narratives against Online Hate Speech: Data and Strategies [Website] [PDF]
Serra Sinem Tekiroğlu, Yi-Ling Chung, and Marco Guerini 8:00–9:00

Recently research has started focusing on avoiding undesired effects that come with content moderation, such as censorship and overblocking, when dealing with hatred online. The core idea is to directly intervene in the discussion with textual responses that are meant to counter the hate content and prevent it from further spreading. Accordingly, automation strategies, such as natural language generation, are beginning to be investigated. Still, they suffer from the lack of sufficient amount of quality data and tend to produce generic/repetitive responses. Being aware of the aforementioned limitations, we present a study on how to collect responses to hate effectively, employing large scale unsupervised language models such as GPT-2 for the generation of silver data, and the best annotation strategies/neural architectures that can be used for data filtering before expert validation/post-editing.

KLEJ: Comprehensive Benchmark for Polish Language Understanding [Website] [PDF]
Piotr Rybak, Robert Mroczkowski, Janusz Tracz, and Ireneusz Gawlik 8:00–9:00

In recent years, a series of Transformer-based models unlocked major improvements in general natural language understanding (NLU) tasks. Such a fast pace of research would not be possible without general NLU benchmarks, which allow for a fair comparison of the proposed methods. However, such benchmarks are available only for a handful of languages. To alleviate this issue, we introduce a comprehensive multi-task benchmark for the Polish language understanding, accompanied by an online leaderboard. It consists of a diverse set of tasks, adopted from existing datasets for named entity recognition, question-answering, textual entailment, and others. We also introduce a new sentiment analysis task for the e-commerce domain, named Allegro Reviews (AR). To ensure a common evaluation scheme and promote models that generalize to different NLU tasks, the benchmark includes datasets from varying domains and applications. Additionally, we release HerBERT, a Transformer-based model trained specifically for the Polish language, which has the best average performance and obtains the best results for three out of nine tasks. Finally, we provide an extensive evaluation, including several standard baselines and recently proposed, multilingual Transformer-based models.

Learning and Evaluating Emotion Lexicons for 91 Languages [Website] [PDF]
Sven Buechel, Susanna Rücker, and Udo Hahn 8:00–9:00

Emotion lexicons describe the affective meaning of words and thus constitute a centerpiece for advanced sentiment and emotion analysis. Yet, manually curated lexicons are only available for a handful of languages, leaving most languages of the world without such a precious resource for downstream applications. Even worse, their coverage is often limited both in terms of the lexical units they contain and the emotional variables they feature. In order to break this bottleneck, we here introduce a methodology for creating almost arbitrarily large emotion lexicons for any target

language. Our approach requires nothing but a source language emotion lexicon, a bilingual word translation model, and a target language embedding model. Fulfilling these requirements for 91 languages, we are able to generate representationally rich high-coverage lexicons comprising eight emotional variables with more than 100k lexical entries each. We evaluated the automatically generated lexicons against human judgment from 26 datasets, spanning 12 typologically diverse languages, and found that our approach produces results in line with state-of-the-art monolingual approaches to lexicon creation and even surpasses human reliability for some languages and variables. Code and data are available at https://github.com/JULIELab/MEmoLon archived under DOI 10.5281/zenodo.3779901.

Multi-Hypothesis Machine Translation Evaluation [Website][PDF]
Marina Fomicheva, Lucia Specia, and Francisco Guzmán 8:00–9:00

Reliably evaluating Machine Translation (MT) through automated metrics is a long-standing problem. One of the main challenges is the fact that multiple outputs can be equally valid. Attempts to minimise this issue include metrics that relax the matching of MT output and reference strings, and the use of multiple references. The latter has been shown to significantly improve the performance of evaluation metrics. However, collecting multiple references is expensive and in practice a single reference is generally used. In this paper, we propose an alternative approach: instead of modelling linguistic variation in human reference we exploit the MT model uncertainty to generate multiple diverse translations and use these: (i) as surrogates to reference translations; (ii) to obtain a quantification of translation variability to either complement existing metric scores or (iii) replace references altogether. We show that for a number of popular evaluation metrics our variability estimates lead to substantial improvements in correlation with human judgements of quality by up 15%.

Multimodal Quality Estimation for Machine Translation [Website][PDF]
Shu Okabe, Frédéric Blain, and Lucia Specia 8:00–9:00

We propose approaches to Quality Estimation (QE) for Machine Translation that explore both text and visual modalities for Multimodal QE. We compare various multimodality integration and fusion strategies. For both sentence-level and document-level predictions, we show that state-of-the-art neural and feature-based QE frameworks obtain better results when using the additional modality.

PuzzLing Machines: A Challenge on Learning From Small Data [Website][PDF]
Gözde Gül Şahin, Yova Kementchedjhieva, Phillip Rust, and Iryna Gurevych 8:00–9:00

Deep neural models have repeatedly proved excellent at memorizing surface patterns from large datasets for various ML and NLP benchmarks. They struggle to achieve human-like thinking, however, because they lack the skill of iterative reasoning upon knowledge. To expose this problem in a new light, we introduce a challenge on learning from small data, PuzzLing Machines, which consists of Rosetta Stone puzzles from Linguistic Olympiads for high school students. These puzzles are carefully designed to contain only the minimal amount of parallel text necessary to deduce the form of unseen expressions. Solving them does not require external information (e.g., knowledge bases, visual signals) or linguistic expertise, but meta-linguistic awareness and deductive skills. Our challenge contains around 100 puzzles covering a wide range of linguistic phenomena from 81 languages. We show that both simple statistical algorithms and state-of-the-art deep neural models perform inadequately on this challenge, as expected. We hope that this benchmark, available at https://ukplab.github.io/PuzzLing-Machines/, inspires further efforts towards a new paradigm in NLP—one that is grounded in human-like reasoning and understanding.

The SOFC-Exp Corpus and Neural Approaches to Information Extraction in the Materials Science Domain [Website][PDF]
Annemarie Friedrich, Heike Adel, Federico Tomazic, Johannes Hingerl, Renou Benteau, Anika Marusczyk, and Lukas Lange 8:00–9:00

This paper presents a new challenging information extraction task in the domain of materials science. We develop an annotation scheme for marking information on experiments related to solid oxide fuel cells in scientific publications, such as involved materials and measurement conditions. With this paper, we publish our annotation guidelines, as well as our SOFC-Exp corpus consisting of 45 open-access scholarly articles annotated by domain experts. A corpus and an inter-annotator agreement study demonstrate the complexity of the suggested named entity recognition and slot filling tasks as well as high annotation quality. We also present strong neural-network based models for a variety of tasks that can be addressed on the basis of our new data set. On all tasks, using BERT embeddings leads to large performance gains, but with increasing task complexity, adding a recurrent neural network on top seems beneficial. Our models will serve as competitive baselines in future work, and analysis of their performance highlights difficult cases when modeling the data and suggests promising research directions.

The TechQA Dataset [Website][PDF]
Vittorio Castelli, Rishav Chakravarti, Saswati Dana, Anthony Ferritto, Radu Florian, Martin Franz, Dinesh Garg, Dinesh Khandelwal, Scott McCarley, Michael McCawley, Mohamed Nasr, Lin Pan, Cezar Pendus, John Pitrelli, Saurabh Pujar, Salim Roukos, Andrzej Sakrajda, Avi Sil, Rosario Uceda-Sosa, Todd Ward, and Rong Zhang 8:00–9:00

We introduce TECHQA, a domain-adaptation question answering dataset for the technical support domain. The TECHQA corpus highlights two real-world issues from the automated customer support domain. First, it contains actual questions posed by users on a technical forum, rather than questions generated specifically for a competition or a task. Second, it has a real-world size — 600 training, 310 dev, and 490 evaluation question/answer pairs — thus reflecting the cost of creating large labeled datasets with actual data. Hence, TECHQA is meant to stimulate research in domain adaptation rather than as a resource to build QA systems from scratch. TECHQA was obtained by crawling the IBMDeveloper and DeveloperWorks forums for questions with accepted answers provided in an IBM Technote—a

technical document that addresses a specific technical issue. We also release a collection of the 801,998 Technotes available on the web as of April 4, 2019 as a companion resource that can be used to learn representations of the IT domain language.

iSarcasm: A Dataset of Intended Sarcasm [Website] [PDF]
Silviu Oprea and Walid Magdy 8:00–9:00

We consider the distinction between intended and perceived sarcasm in the context of textual sarcasm detection. The former occurs when an utterance is sarcastic from the perspective of its author, while the latter occurs when the utterance is interpreted as sarcastic by the audience. We show the limitations of previous labelling methods in capturing intended sarcasm and introduce the iSarcasm dataset of tweets labeled for sarcasm directly by their authors. Examining the state-of-the-art sarcasm detection models on our dataset showed low performance compared to previously studied datasets, which indicates that these datasets might be biased or obvious and sarcasm could be a phenomenon under-studied computationally thus far. By providing the iSarcasm dataset, we aim to encourage future NLP research to develop methods for detecting sarcasm in text as intended by the authors of the text, not as labeled under assumptions that we demonstrate to be sub-optimal.

Session 2A Semantics: Sentence Level-1

AMR Parsing via Graph-Sequence Iterative Inference [Website] [PDF]
Deng Cai and Wai Lam 8:00–9:00

We propose a new end-to-end model that treats AMR parsing as a series of dual decisions on the input sequence and the incrementally constructed graph. At each time step, our model performs multiple rounds of attention, reasoning, and composition that aim to answer two critical questions: (1) which part of the input *sequence* to abstract; and (2) where in the output *graph* to construct the new concept. We show that the answers to these two questions are mutually causalities. We design a model based on iterative inference that helps achieve better answers in both perspectives, leading to greatly improved parsing accuracy. Our experimental results significantly outperform all previously reported SMATCH scores by large margins. Remarkably, without the help of any large-scale pre-trained language model (e.g., BERT), our model already surpasses previous state-of-the-art using BERT. With the help of BERT, we can push the state-of-the-art results to 80.2% on LDC2017T10 (AMR 2.0) and 75.4% on LDC2014T12 (AMR 1.0).

Session 2A: Student Research Workshop

Topic Balancing with Additive Regularization of Topic Models [Website][PDF]
Eugeniia Veselova and Konstantin Vorontsov 8:00–9:00

This article proposes a new approach for building topic models on unbalanced collections in topic modelling, based on the existing methods and our experiments with such methods. Real-world data collections contain topics in various proportions, and often documents of the relatively small theme become distributed all over the larger topics instead of being grouped into one topic. To address this issue, we design a new regularizer for Theta and Phi matrices in probabilistic Latent Semantic Analysis (pLSA) model. We make sure this regularizer increases the quality of topic models, trained on unbalanced collections. Besides, we conceptually support this regularizer by our experiments.

Combining Subword Representations into Word-level Representations in the Transformer Architecture [Website][PDF]
Noe Casas, Marta R. Costa-jussà, and José A. R. Fonollosa 8:00–9:00

In Neural Machine Translation, using word-level tokens leads to degradation in translation quality. The dominant approaches use subword-level tokens, but this increases the length of the sequences and makes it difficult to profit from word-level information such as POS tags or semantic dependencies. We propose a modification to the Transformer model to combine subword-level representations into word-level ones in the first layers of the encoder, reducing the effective length of the sequences in the following layers and providing a natural point to incorporate extra word-level information. Our experiments show that this approach maintains the translation quality with respect to the normal Transformer model when no extra word-level information is injected and that it is superior to the currently dominant method for incorporating word-level source language information to models based on subword-level vocabularies.

Zero-shot North Korean to English Neural Machine Translation by Character Tokenization and Phoneme Decomposition [Website][PDF]
Hwichan Kim, Tosho Hirasawa, and Mamoru Komachi 8:00–9:00

The primary limitation of North Korean to English translation is the lack of a parallel corpus; therefore, high translation accuracy cannot be achieved. To address this problem, we propose a zero-shot approach using South Korean data, which are remarkably similar to North Korean data. We train a neural machine translation model after tokenizing a South Korean text at the character level and decomposing characters into phonemes.We demonstrate that our method can effectively learn North Korean to English translation and improve the BLEU scores by +1.01 points in comparison with the baseline.

Media Bias, the Social Sciences, and NLP: Automating Frame Analyses to Identify Bias by Word Choice and Labeling [Website][PDF]
Felix Hamborg 8:00–9:00

Media bias can strongly impact the public perception of topics reported in the news. A difficult to detect, yet powerful form of slanted news coverage is called bias by word choice and labeling (WCL). WCL bias can occur, for example, when journalists refer to the same semantic concept by using different terms that frame the concept differently and consequently may lead to different assessments by readers, such as the terms “freedom fighters” and “terrorists,” or “gun rights” and “gun control.” In this research project, I aim to devise methods that identify instances of WCL bias and estimate the frames they induce, e.g., not only is “terrorists” of negative polarity but also ascribes to aggression and fear. To achieve this, I plan to research methods using natural language processing and deep learning while employing models and using analysis concepts from the social sciences, where researchers have studied media bias for decades. The first results indicate the effectiveness of this interdisciplinary research approach. My vision is to devise a system that helps news readers to become aware of the differences in media coverage caused by bias.

Session 2A: Summarization-1

A Large-Scale Multi-Document Summarization Dataset from the Wikipedia Current Events Portal [Website][PDF]
Demian Gholipour Ghalandari, Chris Hokamp, Nghia The Pham, John Glover, and Georgiana Ifrim 8:00–9:00

Multi-document summarization (MDS) aims to compress the content in large document collections into short summaries and has important applications in story clustering for newsfeeds, presentation of search results, and timeline generation. However, there is a lack of datasets that realistically address such use cases at a scale large enough for training supervised models for this task. This work presents a new dataset for MDS that is large both in the total number of document clusters and in the size of individual clusters. We build this dataset by leveraging the Wikipedia Current Events Portal (WCEP), which provides concise and neutral human-written summaries of news events, with links to external source articles. We also automatically extend these source articles by looking for related articles in the Common Crawl archive. We provide a quantitative analysis of the dataset and empirical results for several state-of-the-art MDS techniques.

Attend, Translate and Summarize: An Efficient Method for Neural Cross-Lingual Summarization [Website][PDF]
Junnan Zhu, Yu Zhou, Jiajun Zhang, and Chengqing Zong 8:00–9:00

Cross-lingual summarization aims at summarizing a document in one language (e.g., Chinese) into another language (e.g., English). In this paper, we propose a novel method inspired by the translation pattern in the process of obtaining a cross-lingual summary. We first attend to some words in the source text, then translate them into the target language, and summarize to get the final summary. Specifically, we first employ the encoder-decoder attention distribution to attend to the source words. Second, we present three strategies to acquire the translation probability, which helps obtain the translation candidates for each source word. Finally, each summary word is generated either from the neural distribution or from the translation candidates of source words. Experimental results on Chinese-to-English and English-to-Chinese summarization tasks have shown that our proposed method can significantly outperform the baselines, achieving comparable performance with the state-of-the-art.

Examining the State-of-the-Art in News Timeline Summarization [Website][PDF]
Demian Gholipour Ghalandari and Georgiana Ifrim 8:00–9:00

Previous work on automatic news timeline summarization (TLS) leaves an unclear picture about how this task can generally be approached and how well it is currently solved. This is mostly due to the focus on individual subtasks, such as date selection and date summarization, and to the previous lack of appropriate evaluation metrics for the full TLS task. In this paper, we compare different TLS strategies using appropriate evaluation frameworks, and propose a simple and effective combination of methods that improves over the stateof-the-art on all tested benchmarks. For a more robust evaluation, we also present a new TLS dataset, which is larger and spans longer time periods than previous datasets.

Improving Truthfulness of Headline Generation [Website][PDF]
Kazuki Matsumaru, Sho Takase, and Naoaki Okazaki 8:00–9:00

Most studies on abstractive summarization report ROUGE scores between system and reference summaries. However, we have a concern about the truthfulness of generated summaries: whether all facts of a generated summary are mentioned in the source text. This paper explores improving the truthfulness in headline generation on two popular datasets. Analyzing headlines generated by the state-of-the-art encoder-decoder model, we show that the model sometimes generates untruthful headlines. We conjecture that one of the reasons lies in untruthful supervision data used for training the model. In order to quantify the truthfulness of article-headline pairs, we consider the textual entailment of whether an article entails its headline. After confirming quite a few untruthful instances in the datasets, this study hypothesizes that removing untruthful instances from the supervision data may remedy the problem of the untruthful behaviors of the model. Building a binary classifier that predicts an entailment relation between an article and its headline, we filter out untruthful instances from the supervision data. Experimental results demonstrate that the headline generation model trained on filtered supervision data shows no clear difference in ROUGE scores but remarkable improvements in automatic and manual evaluations of the generated headlines.

SUPERT: Towards New Frontiers in Unsupervised Evaluation Metrics for Multi-Document Summarization [Website][PDF]
Yang Gao, Wei Zhao, and Steffen Eger 8:00–9:00

We study unsupervised multi-document summarization evaluation metrics, which require neither human-written reference summaries nor human annotations (e.g. preferences, ratings, etc.). We propose SUPERT, which rates the quality of a summary by measuring its semantic similarity with a pseudo reference summary, i.e. selected salient sentences from the source documents, using contextualized embeddings and soft token alignment techniques. Compared to the state-of-the-art unsupervised evaluation metrics, SUPERT correlates better with human ratings by 18-39%. Furthermore, we use SUPERT as rewards to guide a neural-based reinforcement learning summarizer, yielding favorable performance compared to the state-of-the-art unsupervised summarizers. All source code is available at https://github.com/yg211/acl20-ref-free-eval.

Self-Attention Guided Copy Mechanism for Abstractive Summarization [Website][PDF]
Song Xu, Haoran Li, Peng Yuan, Youzheng Wu, Xiaodong He, and Bowen Zhou 8:00–9:00

Copy module has been widely equipped in the recent abstractive summarization models, which facilitates the decoder to extract words from the source into the summary. Generally, the encoder-decoder attention is served as the copy distribution, while how to guarantee that important words in the source are copied remains a challenge. In this work, we propose a Transformer-based model to enhance the copy mechanism. Specifically, we identify the importance of each source word based on the degree centrality with a directed graph built by the self-attention layer in the Transformer. We use the centrality of each source word to guide the copy process explicitly. Experimental results show that the self-attention graph provides useful guidance for the copy distribution. Our proposed models significantly outperform the baseline methods on the CNN/Daily Mail dataset and the Gigaword dataset.

Demo Session 2B

Time: *8:45–9:30*

Personalized PageRank with Syntagmatic Information for Multilingual Word Sense Disambiguation
[Website] [PDF]
Federico Scozzafava, Marco Maru, Fabrizio Brignone, Giovanni Torrisi, and Roberto Navigli

Exploiting syntagmatic information is an encouraging research focus to be pursued in an effort to close the gap between knowledge-based and supervised Word Sense Disambiguation (WSD) performance. We follow this direction in our next-generation knowledge-based WSD system, SyntagRank, which we make available via a Web interface and a RESTful API. SyntagRank leverages the disambiguated pairs of co-occurring words included in SyntagNet, a lexical-semantic combination resource, to perform state-of-the-art knowledge-based WSD in a multilingual setting. Our service provides both a user-friendly interface, available at http://syntagnet.org/, and a RESTful endpoint to query the system programmatically (accessible at http://api.syntagnet.org/).

Session 2B Overview – Monday, July 6, 2020 9:00–10:00

Track A *Cognitive Modeling and Psycholinguistics-2* Abstracts	[TACL] How Furiously Can Colourless Green Ideas Sleep? Sentence Acceptability in Context *Lau, Armendariz, Purver, Shu, and Lappin* [Website][PDF]	Predicting Depression in Screening Interviews from Latent Categorization of Interview Prompts *Rinaldi, Fox Tree, and Chaturvedi* [Website][PDF]			
Track B *Dialogue and Interactive Systems-4* Abstracts	Beyond User Self-Reported Likert Scale Ratings: A Comparison Model for Automatic Dialog Evaluation *Liang, Zou, and Yu* [Website][PDF]	Conversational Word Embedding for Retrieval-Based Dialog System *Ma, Cui, Liu, Wang, Wang, and Hu* [Website][PDF]	Designing Precise and Robust Dialogue Response Evaluators *Zhao, Lala, and Kawahara* [Website][PDF]	Evaluating Dialogue Generation Systems via Response Selection *Sato, Akama, Ouchi, Suzuki, and Inui* [Website][PDF]	Few-shot Slot Tagging with Collapsed Dependency Transfer and Label-enhanced Task-adaptive Projection Network *Hou, Che, Lai, Zhou, Liu, Liu, and Liu* [Website][PDF]
	Generating Informative Conversational Response using Recurrent Knowledge-Interaction and Knowledge-Copy *Lin, Jian, He, Wang, and Chu* [Website][PDF]	Guiding Variational Response Generator to Exploit Persona *Wu, Li, Wang, Chen, Wong, Huang, and Wang* [Website][PDF]	Learning Dialog Policies from Weak Demonstrations *Gordon-Hall, Gorinski, and Cohen* [Website][PDF]	MuTual: A Dataset for Multi-Turn Dialogue Reasoning *Cui, Wu, Liu, Zhang, and Zhou* [Website][PDF]	You Impress Me: Dialogue Generation via Mutual Persona Perception *Liu, Chen, Chen, LOU, Chen, Zhou, and Zhang* [Website][PDF]
Track C *Discourse and Pragmatics-2* Abstracts	Bridging Anaphora Resolution as Question Answering *Hou* [Website][PDF]	Dialogue Coherence Assessment Without Explicit Dialogue Act Labels *Mesgar, Bücker, and Gurevych* [Website][PDF]			
Track D *Generation-4* Abstracts	Explicit Semantic Decomposition for Definition Generation *Li, Bao, Huang, Dai, and CHEN* [Website][PDF]	Fast and Accurate Non-Projective Dependency Tree Linearization *Yu, Tannert, Vu, and Kuhn* [Website][PDF]	Semantic Graphs for Generating Deep Questions *Pan, Xie, Feng, Chua, and Kan* [Website][PDF]	Syn-QG: Syntactic and Shallow Semantic Rules for Question Generation *Dhole and Manning* [Website][PDF]	Unsupervised Paraphrasing by Simulated Annealing *Liu, Mou, Meng, Zhou, Zhou, and Song* [Website][PDF]
Track E *Information Extraction-1* Abstracts	A Novel Cascade Binary Tagging Framework for Relational Triple Extraction *Wei, Su, Wang, Tian, and Chang* [Website][PDF]	In Layman's Terms: Semi-Open Relation Extraction from Scientific Texts *Kruiper, Vincent, Chen-Burger, Desmulliez, and Konstas* [Website][PDF]	NAT: Noise-Aware Training for Robust Neural Sequence Labeling *Namysl, Behnke, and Köhler* [Website][PDF]	Named Entity Recognition without Labelled Data: A Weak Supervision Approach *Lison, Barnes, Hubin, and Touileb* [Website][PDF]	Probing Linguistic Features of Sentence-Level Representations in Relation Extraction *Alt, Gabryszak, and Hennig* [Website][PDF]

	Reasoning with Latent Structure Refinement for Document-Level Relation Extraction *Nan, Guo, Sekulic, and Lu* [Website][PDF]	TACRED Revisited: A Thorough Evaluation of the TACRED Relation Extraction Task *Alt, Gabryszak, and Hennig* [Website][PDF]			
Track F *Machine Translation-2* Abstracts	Bilingual Dictionary Based Neural Machine Translation without Using Parallel Sentences *Duan, Ji, Jia, Tan, Zhang, Chen, Luo, and Zhang* [Website][PDF]	Boosting Neural Machine Translation with Similar Translations *XU, Crego, and Senellart* [Website][PDF]	Character-Level Translation with Self-attention *Gao, Nikolov, Hu, and Hahnloser* [Website][PDF]	End-to-End Neural Word Alignment Outperforms GIZA++ *Zenkel, Wuebker, and DeNero* [Website][PDF]	Enhancing Machine Translation with Dependency-Aware Self-Attention *Bugliarello and Okazaki* [Website][PDF]
	Improving Massively Multilingual Neural Machine Translation and Zero-Shot Translation *Zhang, Williams, Titov, and Sennrich* [Website][PDF]	It's Easier to Translate out of English than into it: Measuring Neural Translation Difficulty by Cross-Mutual Information *Bugliarello, Mielke, Anastasopoulos, Cotterell, and Okazaki* [Website][PDF]	Language-aware Interlingua for Multilingual Neural Machine Translation *Zhu, Yu, Cheng, and Luo* [Website][PDF]	Norm-Based Curriculum Learning for Neural Machine Translation *Liu, Lai, Wong, and Chao* [Website][PDF]	On the Limitations of Cross-lingual Encoders as Exposed by Reference-Free Machine Translation Evaluation *Zhao, Glavaš, Peyrard, Gao, West, and Eger* [Website][PDF]
	Parallel Sentence Mining by Constrained Decoding *Chen, Bogoychev, Heafield, and Kirefu* [Website][PDF]	Self-Attention with Cross-Lingual Position Representation *Ding, Wang, and Tao* [Website][PDF]	"You Sound Just Like Your Father" Commercial Machine Translation Systems Include Stylistic Biases *Hovy, Bianchi, and Fornaciari* [Website][PDF]		
Track G *NLP Applications-2* Abstracts	Fast and Accurate Deep Bidirectional Language Representations for Unsupervised Learning *Shin, Lee, Yoon, and Jung* [Website][PDF]	Fine-grained Interest Matching for Neural News Recommendation *Wang, Wu, Liu, and Xie* [Website][PDF]	Interpretable Operational Risk Classification with Semi-Supervised Variational Autoencoder *Zhou, Zhang, and Yang* [Website][PDF]	Interpreting Twitter User Geolocation *Zhong, Wang, Zhou, Trajcevski, Zhang, and Yang* [Website][PDF]	MMPE: A Multi-Modal Interface for Post-Editing Machine Translation *Herbig, Düwel, Pal, Meladaki, Monshizadeh, Krüger, and Genabith* [Website][PDF]
	Modeling Code-Switch Languages Using Bilingual Parallel Corpus *Lee and Li* [Website][PDF]	SpellGCN: Incorporating Phonological and Visual Similarities into Language Models for Chinese Spelling Check *Cheng, Xu, Chen, Jiang, Wang, Wang, Chu, and Qi* [Website][PDF]	Spelling Error Correction with Soft-Masked BERT *Zhang, Huang, Liu, and Li* [Website][PDF]		

Track H *Resources and Evaluation-3* Abstracts	A Diverse Corpus for Evaluating and Developing English Math Word Problem Solvers *Miao, Liang, and Su* [Website][PDF]	A Monolingual Approach to Contextualized Word Embeddings for Mid-Resource Languages *Ortiz Suárez, Romary, and Sagot* [Website][PDF]	Improving Image Captioning Evaluation by Considering Inter References Variance *Yi, Deng, and Hu* [Website][PDF]	Revisiting the Context Window for Cross-lingual Word Embeddings *Ri and Tsuruoka* [Website][PDF]	Will-They-Won't-They: A Very Large Dataset for Stance Detection on Twitter *Conforti, Berndt, Pilehvar, Giannitsarou, Toxvaerd, and Collier* [Website][PDF]
Track I *Student Research Workshop* Abstracts	SCAR: Sentence Compression using Autoencoders for Reconstruction *Malireddy, Maniar, and Shrivastava* [Website][PDF]	Feature Difference Makes Sense: A medical image captioning model exploiting feature difference and tag information *Park, Kim, Yoon, Park, and Choi* [Website][PDF]	Multi-Task Neural Model for Agglutinative Language Translation *Pan, Li, Yang, and Dong* [Website][PDF]	Considering Likelihood in NLP Classification Explanations with Occlusion and Language Modeling *Harbecke and Alt* [Website][PDF]	
Track J *Theory and Formalism in NLP (Linguistic and Mathematical)-2* Abstracts	A Three-Parameter Rank-Frequency Relation in Natural Languages *Ding, Utiyama, and Sumita* [Website][PDF]	Dice Loss for Data-imbalanced NLP Tasks *Li, Sun, Meng, Liang, Wu, and Li* [Website][PDF]	Language Models as an Alternative Evaluator of Word Order Hypotheses: A Case Study in Japanese *Kuribayashi, Ito, Suzuki, and Inui* [Website][PDF]	[TACL] Theoretical Limitations of Self-Attention in Neural Sequence Models *Hahn* [Website][PDF]	

Session 2B Details

Session 2B: Cognitive Modeling and Psycholinguistics-2

[TACL] How Furiously Can Colourless Green Ideas Sleep? Sentence Acceptability in Context [Website][PDF]
Jey Han Lau, Carlos Santos Armendariz, Matthew Purver, Chang Shu, and Shalom Lappin 9:00–10:00

We study the influence of context on sentence acceptability. First we compare the acceptability ratings of sentences judged in isolation, with a relevant context, and with an irrelevant context. Our results show that context induces a cognitive load for humans, which compresses the distribution of ratings. Moreover, in relevant contexts we observe a discourse coherence effect which uniformly raises acceptability. Next, we test unidirectional and bidirectional language models in their ability to predict acceptability ratings. The bidirectional models show very promising results, with the best model achieving a new state-of-the-art for unsupervised acceptability prediction. The two sets of experiments provide insights into the cognitive aspects of sentence processing and central issues in the computational modelling of text and discourse.

Predicting Depression in Screening Interviews from Latent Categorization of Interview Prompts [Website][PDF]
Alex Rinaldi, Jean Fox Tree, and Snigdha Chaturvedi 9:00–10:00

Accurately diagnosing depression is difficult– requiring time-intensive interviews, assessments, and analysis. Hence, automated methods that can assess linguistic patterns in these interviews could help psychiatric professionals make faster, more informed decisions about diagnosis. We propose JLPC, a model that analyzes interview transcripts to identify depression while jointly categorizing interview prompts into latent categories. This latent categorization allows the model to define high-level conversational contexts that influence patterns of language in depressed individuals. We show that the proposed model not only outperforms competitive baselines, but that its latent prompt categories provide psycholinguistic insights about depression.

Session 2B: Dialogue and Interactive Systems-4

Beyond User Self-Reported Likert Scale Ratings: A Comparison Model for Automatic Dialog Evaluation [Website][PDF]
Weixin Liang, James Zou, and Zhou Yu 9:00–10:00

Open Domain dialog system evaluation is one of the most important challenges in dialog research. Existing automatic evaluation metrics, such as BLEU are mostly reference-based. They calculate the difference between the generated response and a limited number of available references. Likert-score based self-reported user rating is widely adopted by social conversational systems, such as Amazon Alexa Prize chatbots. However, self-reported user rating suffers from bias and variance among different users. To alleviate this problem, we formulate dialog evaluation as a comparison task. We also propose an automatic evaluation model CMADE (Comparison Model for Automatic Dialog Evaluation) that automatically cleans self-reported user ratings as it trains on them. Specifically, we first use a self-supervised method to learn better dialog feature representation, and then use KNN and Shapley to remove confusing samples. Our experiments show that CMADE achieves 89.2% accuracy in the dialog comparison task.

Conversational Word Embedding for Retrieval-Based Dialog System [Website][PDF]
Wentao Ma, Yiming Cui, Ting Liu, Dong Wang, Shijin Wang, and Guoping Hu 9:00–10:00

Human conversations contain many types of information, e.g., knowledge, common sense, and language habits. In this paper, we propose a conversational word embedding method named PR-Embedding, which utilizes the conversation pairs <post, reply> to learn word embedding. Different from previous works, PR-Embedding uses the vectors from two different semantic spaces to represent the words in post and reply.To catch the information among the pair, we first introduce the word alignment model from statistical machine translation to generate the cross-sentence window, then train the embedding on word-level and sentence-level.We evaluate the method on single-turn and multi-turn response selection tasks for retrieval-based dialog systems.The experiment results show that PR-Embedding can improve the quality of the selected response.

Designing Precise and Robust Dialogue Response Evaluators [Website][PDF]
Tianyu Zhao, Divesh Lala, and Tatsuya Kawahara 9:00–10:00

Automatic dialogue response evaluator has been proposed as an alternative to automated metrics and human evaluation. However, existing automatic evaluators achieve only moderate correlation with human judgement and they are not robust. In this work, we propose to build a reference-free evaluator and exploit the power of semi-supervised training and pretrained (masked) language models. Experimental results demonstrate that the proposed evaluator achieves a strong correlation (> 0.6) with human judgement and generalizes robustly to diverse responses and corpora. We open-source the code and data in https://github.com/ZHAOTING/dialog-processing.

Evaluating Dialogue Generation Systems via Response Selection [Website][PDF]
Shiki Sato, Reina Akama, Hiroki Ouchi, Jun Suzuki, and Kentaro Inui 9:00–10:00

Existing automatic evaluation metrics for open-domain dialogue response generation systems correlate poorly with human evaluation. We focus on evaluating response generation systems via response selection. To evaluate systems properly via response selection, we propose a method to construct response selection test sets with well-chosen false candidates. Specifically, we propose to construct test sets filtering out some types of false candidates: (i) those unrelated to the ground-truth response and (ii) those acceptable as appropriate responses. Through experiments, we demonstrate that evaluating systems via response selection with the test set developed by our method correlates more strongly with human evaluation, compared with widely used automatic evaluation metrics such as BLEU.

Few-shot Slot Tagging with Collapsed Dependency Transfer and Label-enhanced Task-adaptive Projection Network [Website][PDF]
Yutai Hou, Wanxiang Che, Yongkui Lai, Zhihan Zhou, Yijia Liu, Han Liu, and Ting Liu 9:00–10:00

In this paper, we explore the slot tagging with only a few labeled support sentences (a.k.a. few-shot). Few-shot slot tagging faces a unique challenge compared to the other fewshot classification problems as it calls for modeling the dependencies between labels. But it is hard to apply previously learned label dependencies to an unseen domain, due to the discrepancy of label sets. To tackle this, we introduce a collapsed dependency transfer mechanism into the conditional random field (CRF) to transfer abstract label dependency patterns as transition scores. In the few-shot setting, the emission score of CRF can be calculated as a word's similarity to the representation of each label. To calculate such similarity, we propose a Label-enhanced Task-Adaptive Projection Network (L-TapNet) based on the state-of-the-art few-shot classification model — TapNet, by leveraging label name semantics in representing labels. Experimental results show that our model significantly outperforms the strongest few-shot learning baseline by 14.64 F1 scores in the one-shot setting.

Generating Informative Conversational Response using Recurrent Knowledge-Interaction and Knowledge-Copy [Website][PDF]
Xiexiong Lin, Weiyu Jian, Jianshan He, Taifeng Wang, and Wei Chu 9:00–10:00

Knowledge-driven conversation approaches have achieved remarkable research attention recently. However, generating an informative response with multiple relevant knowledge without losing fluency and coherence is still one of the main challenges. To address this issue, this paper proposes a method that uses recurrent knowledge interaction among response decoding steps to incorporate appropriate knowledge. Furthermore, we introduce a knowledge copy mechanism using a knowledge-aware pointer network to copy words from external knowledge according to knowledge attention distribution. Our joint neural conversation model which integrates recurrent Knowledge-Interaction

and knowledge Copy (KIC) performs well on generating informative responses. Experiments demonstrate that our model with fewer parameters yields significant improvements over competitive baselines on two datasets Wizard-of-Wikipedia(average Bleu +87%; abs.: 0.034) and DuConv(average Bleu +20%; abs.: 0.047)) with different knowledge formats (textual & structured) and different languages (English & Chinese).

Guiding Variational Response Generator to Exploit Persona [Website][PDF]
Bowen Wu, MengYuan Li, Zongsheng Wang, Yifu Chen, Derek F. Wong, qihang feng qihang, Junhong Huang, and Baoxun Wang 9:00–10:00

Leveraging persona information of users in Neural Response Generators (NRG) to perform personalized conversations has been considered as an attractive and important topic in the research of conversational agents over the past few years. Despite of the promising progress achieved by recent studies in this field, persona information tends to be incorporated into neural networks in the form of user embeddings, with the expectation that the persona can be involved via End-to-End learning. This paper proposes to adopt the personality-related characteristics of human conversations into variational response generators, by designing a specific conditional variational autoencoder based deep model with two new regularization terms employed to the loss function, so as to guide the optimization towards the direction of generating both persona-aware and relevant responses. Besides, to reasonably evaluate the performances of various persona modeling approaches, this paper further presents three direct persona-oriented metrics from different perspectives. The experimental results have shown that our proposed methodology can notably improve the performance of persona-aware response generation, and the metrics are reasonable to evaluate the results.

Learning Dialog Policies from Weak Demonstrations [Website][PDF]
Gabriel Gordon-Hall, Philip John Gorinski, and Shay B. Cohen 9:00–10:00

Deep reinforcement learning is a promising approach to training a dialog manager, but current methods struggle with the large state and action spaces of multi-domain dialog systems. Building upon Deep Q-learning from Demonstrations (DQfD), an algorithm that scores highly in difficult Atari games, we leverage dialog data to guide the agent to successfully respond to a user's requests. We make progressively fewer assumptions about the data needed, using labeled, reduced-labeled, and even unlabeled data to train expert demonstrators. We introduce Reinforced Fine-tune Learning, an extension to DQfD, enabling us to overcome the domain gap between the datasets and the environment. Experiments in a challenging multi-domain dialog system framework validate our approaches, and get high success rates even when trained on out-of-domain data.

MuTual: A Dataset for Multi-Turn Dialogue Reasoning [Website][PDF]
Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou 9:00–10:00

Non-task oriented dialogue systems have achieved great success in recent years due to largely accessible conversation data and the development of deep learning techniques. Given a context, current systems are able to yield a relevant and fluent response, but sometimes make logical mistakes because of weak reasoning capabilities. To facilitate the conversation reasoning research, we introduce MuTual, a novel dataset for Multi-Turn dialogue Reasoning, consisting of 8,860 manually annotated dialogues based on Chinese student English listening comprehension exams. Compared to previous benchmarks for non-task oriented dialogue systems, MuTual is much more challenging since it requires a model that be able to handle various reasoning problems. Empirical results show that state-of-the-art methods only reach 71%, which is far behind human performance of 94%, indicating that there is ample room for improving reasoning ability.

You Impress Me: Dialogue Generation via Mutual Persona Perception [Website][PDF]
Qian Liu, Yihong Chen, Bei Chen, Jian-Guang LOU, Zixuan Chen, Bin Zhou, and Dongmei Zhang 9:00–10:00

Despite the continuing efforts to improve the engagingness and consistency of chit-chat dialogue systems, the majority of current work simply focus on mimicking human-like responses, leaving understudied the aspects of modeling understanding between interlocutors. The research in cognitive science, instead, suggests that understanding is an essential signal for a high-quality chit-chat conversation. Motivated by this, we propose Pˆ2 Bot, a transmitter-receiver based framework with the aim of explicitly modeling understanding. Specifically, Pˆ2 Bot incorporates mutual persona perception to enhance the quality of personalized dialogue generation. Experiments on a large public dataset, Persona-Chat, demonstrate the effectiveness of our approach, with a considerable boost over the state-of-the-art baselines across both automatic metrics and human evaluations.

Session 2B: Discourse and Pragmatics-2

Bridging Anaphora Resolution as Question Answering [Website][PDF]
Yufang Hou 9:00–10:00

Most previous studies on bridging anaphora resolution (Poesio et al., 2004; Hou et al., 2013b; Hou, 2018a) use the pairwise model to tackle the problem and assume that the gold mention information is given. In this paper, we cast bridging anaphora resolution as question answering based on context. This allows us to find the antecedent for a given anaphor without knowing any gold mention information (except the anaphor itself). We present a question answering framework (BARQA) for this task, which leverages the power of transfer learning. Furthermore, we propose a novel method to generate a large amount of "quasi-bridging" training data. We show that our model pre-trained on this dataset and fine-tuned on a small amount of in-domain dataset achieves new state-of-the-art results for bridging anaphora resolution on two bridging corpora (ISNotes (Markert et al., 2012) and BASHI (Rosiger, 2018)).

Dialogue Coherence Assessment Without Explicit Dialogue Act Labels [Website][PDF]
Mohsen Mesgar, Sebastian Bücker, and Iryna Gurevych 9:00–10:00

Recent dialogue coherence models use the coherence features designed for monologue texts, e.g. nominal entities, to represent utterances and then explicitly augment them with dialogue-relevant features, e.g., dialogue act labels. It indicates two drawbacks, (a) semantics of utterances are limited to entity mentions, and (b) the performance of coherence models strongly relies on the quality of the input dialogue act labels. We address these issues by introducing a novel approach to dialogue coherence assessment. We use dialogue act prediction as an auxiliary task in a multi-task learning scenario to obtain informative utterance representations for coherence assessment. Our approach alleviates the need for explicit dialogue act labels during evaluation. The results of our experiments show that our model substantially (more than 20 accuracy points) outperforms its strong competitors on the DailyDialogue corpus, and performs on par with them on the SwitchBoard corpus for ranking dialogues concerning their coherence. We release our source code.

Session 2B: Generation-4

Explicit Semantic Decomposition for Definition Generation [Website][PDF]
Jiahuan Li, Yu Bao, Shujian Huang, Xinyu Dai, and Jiajun CHEN 9:00–10:00

Definition generation, which aims to automatically generate dictionary definitions for words, has recently been proposed to assist the construction of dictionaries and help people understand unfamiliar texts. However, previous works hardly consider explicitly modeling the "components" of definitions, leading to under-specific generation results. In this paper, we propose ESD, namely Explicit Semantic Decomposition for definition Generation, which explicitly decomposes the meaning of words into semantic components, and models them with discrete latent variables for definition generation. Experimental results show that ESD achieves top results on WordNet and Oxford benchmarks, outperforming strong previous baselines.

Fast and Accurate Non-Projective Dependency Tree Linearization [Website][PDF]
Xiang Yu, Simon Tannert, Ngoc Thang Vu, and Jonas Kuhn 9:00–10:00

We propose a graph-based method to tackle the dependency tree linearization task. We formulate the task as a Traveling Salesman Problem (TSP), and use a biaffine attention model to calculate the edge costs. We facilitate the decoding by solving the TSP for each subtree and combining the solution into a projective tree. We then design a transition system as post-processing, inspired by non-projective transition-based parsing, to obtain non-projective sentences. Our proposed method outperforms the state-of-the-art linearizer while being 10 times faster in training and decoding.

Semantic Graphs for Generating Deep Questions [Website][PDF]
Liangming Pan, Yuxi Xie, Yansong Feng, Tat-Seng Chua, and Min-Yen Kan 9:00–10:00

This paper proposes the problem of Deep Question Generation (DQG), which aims to generate complex questions that require reasoning over multiple pieces of information about the input passage. In order to capture the global structure of the document and facilitate reasoning, we propose a novel framework that first constructs a semantic-level graph for the input document and then encodes the semantic graph by introducing an attention-based GGNN (Att-GGNN). Afterward, we fuse the document-level and graph-level representations to perform joint training of content selection and question decoding. On the HotpotQA deep-question centric dataset, our model greatly improves performance over questions requiring reasoning over multiple facts, leading to state-of-the-art performance. The code is publicly available at https://github.com/WING-NUS/SG-Deep-Question-Generation.

Syn-QG: Syntactic and Shallow Semantic Rules for Question Generation [Website][PDF]
Kaustubh Dhole and Christopher D. Manning 9:00–10:00

Question Generation (QG) is fundamentally a simple syntactic transformation; however, many aspects of semantics influence what questions are good to form. We implement this observation by developing Syn-QG, a set of transparent syntactic rules leveraging universal dependencies, shallow semantic parsing, lexical resources, and custom rules which transform declarative sentences into question-answer pairs. We utilize PropBank argument descriptions and VerbNet state predicates to incorporate shallow semantic content, which helps generate questions of a descriptive nature and produce inferential and semantically richer questions than existing systems. In order to improve syntactic fluency and eliminate grammatically incorrect questions, we employ back-translation over the output of these syntactic rules. A set of crowd-sourced evaluations shows that our system can generate a larger number of highly grammatical and relevant questions than previous QG systems and that back-translation drastically improves grammaticality at a slight cost of generating irrelevant questions.

Unsupervised Paraphrasing by Simulated Annealing [Website][PDF]
Xianggen Liu, Lili Mou, Fandong Meng, Hao Zhou, Jie Zhou, and Sen Song 9:00–10:00

We propose UPSA, a novel approach that accomplishes Unsupervised Paraphrasing by Simulated Annealing. We model paraphrase generation as an optimization problem and propose a sophisticated objective function, involving semantic similarity, expression diversity, and language fluency of paraphrases. UPSA searches the sentence space towards this objective by performing a sequence of local editing. We evaluate our approach on various datasets, namely, Quora, Wikianswers, MSCOCO, and Twitter. Extensive results show that UPSA achieves the state-of-the-art performance compared with previous unsupervised methods in terms of both automatic and human evaluations. Further, our approach outperforms most existing domain-adapted supervised models, showing the generalizability of UPSA.

Session 2B: Information Extraction-1

A Novel Cascade Binary Tagging Framework for Relational Triple Extraction [Website][PDF]
Zhepei Wei, Jianlin Su, Yue Wang, Yuan Tian, and Yi Chang 9:00–10:00

Extracting relational triples from unstructured text is crucial for large-scale knowledge graph construction. However, few existing works excel in solving the overlapping triple problem where multiple relational triples in the same sentence share the same entities. In this work, we introduce a fresh perspective to revisit the relational triple extraction task and propose a novel cascade binary tagging framework (CasRel) derived from a principled problem formulation. Instead of treating relations as discrete labels as in previous works, our new framework models relations as functions that map subjects to objects in a sentence, which naturally handles the overlapping problem. Experiments show that the CasRel framework already outperforms state-of-the-art methods even when its encoder module uses a randomly initialized BERT encoder, showing the power of the new tagging framework. It enjoys further performance boost when employing a pre-trained BERT encoder, outperforming the strongest baseline by 17.5 and 30.2 absolute gain in F1-score on two public datasets NYT and WebNLG, respectively. In-depth analysis on different scenarios of overlapping triples shows that the method delivers consistent performance gain across all these scenarios. The source code and data are released online.

In Layman's Terms: Semi-Open Relation Extraction from Scientific Texts [Website][PDF]
Ruben Kruiper, Julian Vincent, Jessica Chen-Burger, Marc Desmulliez, and Ioannis Konstas 9:00–10:00

Information Extraction (IE) from scientific texts can be used to guide readers to the central information in scientific documents. But narrow IE systems extract only a fraction of the information captured, and Open IE systems do not perform well on the long and complex sentences encountered in scientific texts. In this work we combine the output of both types of systems to achieve Semi-Open Relation Extraction, a new task that we explore in the Biology domain. First, we present the Focused Open Biological Information Extraction (FOBIE) dataset and use FOBIE to train a state-of-the-art narrow scientific IE system to extract trade-off relations and arguments that are central to biology texts. We then run both the narrow IE system and a state-of-the-art Open IE system on a corpus of 10K open-access scientific biological texts. We show that a significant amount (65%) of erroneous and uninformative Open IE extractions can be filtered using narrow IE extractions. Furthermore, we show that the retained extractions are significantly more often informative to a reader.

NAT: Noise-Aware Training for Robust Neural Sequence Labeling [Website][PDF]
Marcin Namysl, Sven Behnke, and Joachim Köhler 9:00–10:00

Sequence labeling systems should perform reliably not only under ideal conditions but also with corrupted inputs—as these systems often process user-generated text or follow an error-prone upstream component. To this end, we formulate the noisy sequence labeling problem, where the input may undergo an unknown noising process and propose two Noise-Aware Training (NAT) objectives that improve robustness of sequence labeling performed on perturbed input: Our data augmentation method trains a neural model using a mixture of clean and noisy samples, whereas our stability training algorithm encourages the model to create a noise-invariant latent representation. We employ a vanilla noise model at training time. For evaluation, we use both the original data and its variants perturbed with real OCR errors and misspellings. Extensive experiments on English and German named entity recognition benchmarks confirmed that NAT consistently improved robustness of popular sequence labeling models, preserving accuracy on the original input. We make our code and data publicly available for the research community.

Named Entity Recognition without Labelled Data: A Weak Supervision Approach [Website][PDF]
Pierre Lison, Jeremy Barnes, Aliaksandr Hubin, and Samia Touileb 9:00–10:00

Named Entity Recognition (NER) performance often degrades rapidly when applied to target domains that differ from the texts observed during training. When in-domain labelled data is available, transfer learning techniques can be used to adapt existing NER models to the target domain. But what should one do when there is no hand-labelled data for the target domain? This paper presents a simple but powerful approach to learn NER models in the absence of labelled data through weak supervision. The approach relies on a broad spectrum of labelling functions to automatically annotate texts from the target domain. These annotations are then merged together using a hidden Markov model which captures the varying accuracies and confusions of the labelling functions. A sequence labelling model can finally be trained on the basis of this unified annotation. We evaluate the approach on two English datasets (CoNLL 2003 and news articles from Reuters and Bloomberg) and demonstrate an improvement of about 7 percentage points in entity-level F1 scores compared to an out-of-domain neural NER model.

Probing Linguistic Features of Sentence-Level Representations in Relation Extraction [Website][PDF]
Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig 9:00–10:00

Despite the recent progress, little is known about the features captured by state-of-the-art neural relation extraction (RE) models. Common methods encode the source sentence, conditioned on the entity mentions, before classifying the relation. However, the complexity of the task makes it difficult to understand how encoder architecture and supporting linguistic knowledge affect the features learned by the encoder. We introduce 14 probing tasks targeting linguistic properties relevant to RE, and we use them to study representations learned by more than 40 different encoder architecture and linguistic feature combinations trained on two datasets, TACRED and SemEval 2010 Task 8. We find that the bias induced by the architecture and the inclusion of linguistic features are clearly expressed in the probing task performance. For example, adding contextualized word representations greatly increases performance on probing tasks with a focus on named entity and part-of-speech information, and yields better results in RE. In contrast, entity masking improves RE, but considerably lowers performance on entity type related probing tasks.

Reasoning with Latent Structure Refinement for Document-Level Relation Extraction [Website][PDF]
Guoshun Nan, Zhijiang Guo, Ivan Sekulic, and Wei Lu 9:00–10:00

Document-level relation extraction requires integrating information within and across multiple sentences of a document and capturing complex interactions between inter-sentence entities. However, effective aggregation of relevant information in the document remains a challenging research question. Existing approaches construct static document-level graphs based on syntactic trees, co-references or heuristics from the unstructured text to model the dependencies. Unlike previous methods that may not be able to capture rich non-local interactions for inference, we propose a novel model that empowers the relational reasoning across sentences by automatically inducing the latent document-level graph. We further develop a refinement strategy, which enables the model to incrementally aggregate relevant information for multi-hop reasoning. Specifically, our model achieves an F1 score of 59.05 on a large-scale document-level dataset (DocRED), significantly improving over the previous results, and also yields new state-of-the-art results on the CDR and GDA dataset. Furthermore, extensive analyses show that the model is able to discover more accurate inter-sentence relations.

TACRED Revisited: A Thorough Evaluation of the TACRED Relation Extraction Task [Website][PDF]
Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig 9:00–10:00

TACRED is one of the largest, most widely used crowdsourced datasets in Relation Extraction (RE). But, even with recent advances in unsupervised pre-training and knowledge enhanced neural RE, models still show a high error rate. In this paper, we investigate the questions: Have we reached a performance ceiling or is there still room for improvement? And how do crowd annotations, dataset, and models contribute to this error rate? To answer these questions, we first validate the most challenging 5K examples in the development and test sets using trained annotators. We find that label errors account for 8% absolute F1 test error, and that more than 50% of the examples need to be relabeled. On the relabeled test set the average F1 score of a large baseline model set improves from 62.1 to 70.1. After validation, we analyze misclassifications on the challenging instances, categorize them into linguistically motivated error groups, and verify the resulting error hypotheses on three state-of-the-art RE models. We show that two groups of ambiguous relations are responsible for most of the remaining errors and that models may adopt shallow heuristics on the dataset when entities are not masked.

Session 2B: Machine Translation-2

Bilingual Dictionary Based Neural Machine Translation without Using Parallel Sentences [Website][PDF]
Xiangyu Duan, Baijun Ji, Hao Jia, Min Tan, Min Zhang, Boxing Chen, Weihua Luo, and Yue Zhang 9:00–10:00

In this paper, we propose a new task of machine translation (MT), which is based on no parallel sentences but can refer to a ground-truth bilingual dictionary. Motivated by the ability of a monolingual speaker learning to translate via looking up the bilingual dictionary, we propose the task to see how much potential an MT system can attain using the bilingual dictionary and large scale monolingual corpora, while is independent on parallel sentences. We propose anchored training (AT) to tackle the task. AT uses the bilingual dictionary to establish anchoring points for closing the gap between source language and target language. Experiments on various language pairs show that our approaches are significantly better than various baselines, including dictionary-based word-by-word translation, dictionary-supervised cross-lingual word embedding transformation, and unsupervised MT. On distant language pairs that are hard for unsupervised MT to perform well, AT performs remarkably better, achieving performances comparable to supervised SMT trained on more than 4M parallel sentences.

Boosting Neural Machine Translation with Similar Translations [Website][PDF]
Jitao XU, Josep Crego, and Jean Senellart 9:00–10:00

This paper explores data augmentation methods for training Neural Machine Translation to make use of similar translations, in a comparable way a human translator employs fuzzy matches. In particular, we show how we can simply present the neural model with information of both source and target sides of the fuzzy matches, we also extend the similarity to include semantically related translations retrieved using sentence distributed representations. We show that translations based on fuzzy matching provide the model with "copy" information while translations based on embedding similarities tend to extend the translation "context". Results indicate that the effect from both similar sentences are adding up to further boost accuracy, combine naturally with model fine-tuning and are providing dynamic adaptation for unseen translation pairs. Tests on multiple data sets and domains show consistent accuracy improvements. To foster research around these techniques, we also release an Open-Source toolkit with efficient and flexible fuzzy-match implementation.

Character-Level Translation with Self-attention [Website][PDF]
Yingqiang Gao, Nikola I. Nikolov, Yuhuang Hu, and Richard H.R. Hahnloser 9:00–10:00

We explore the suitability of self-attention models for character-level neural machine translation. We test the standard transformer model, as well as a novel variant in which the encoder block combines information from nearby characters using convolutions. We perform extensive experiments on WMT and UN datasets, testing both bilingual and multilingual translation to English using up to three input languages (French, Spanish, and Chinese). Our transformer variant consistently outperforms the standard transformer at the character-level and converges faster while learning more robust character-level alignments.

End-to-End Neural Word Alignment Outperforms GIZA++ [Website][PDF]
Thomas Zenkel, Joern Wuebker, and John DeNero 9:00–10:00

Word alignment was once a core unsupervised learning task in natural language processing because of its essential role in training statistical machine translation (MT) models. Although unnecessary for training neural MT models, word alignment still plays an important role in interactive applications of neural machine translation, such as annotation transfer and lexicon injection. While statistical MT methods have been replaced by neural approaches with superior performance, the twenty-year-old GIZA++ toolkit remains a key component of state-of-the-art word alignment systems. Prior work on neural word alignment has only been able to outperform GIZA++ by using its output during training. We present the first end-to-end neural word alignment method that consistently outperforms GIZA++ on three data sets. Our approach repurposes a Transformer model trained for supervised translation to also serve as an unsupervised word alignment model in a manner that is tightly integrated and does not affect translation quality.

Enhancing Machine Translation with Dependency-Aware Self-Attention [Website][PDF]
Emanuele Bugliarello and Naoaki Okazaki 9:00–10:00

Most neural machine translation models only rely on pairs of parallel sentences, assuming syntactic information is automatically learned by an attention mechanism. In this work, we investigate different approaches to incorporate syntactic knowledge in the Transformer model and also propose a novel, parameter-free, dependency-aware self-attention mechanism that improves its translation quality, especially for long sentences and in low-resource scenarios. We show the efficacy of each approach on WMT English-German and English-Turkish, and WAT English-Japanese translation tasks.

Improving Massively Multilingual Neural Machine Translation and Zero-Shot Translation [Website][PDF]
Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich 9:00–10:00

Massively multilingual models for neural machine translation (NMT) are theoretically attractive, but often underperform bilingual models and deliver poor zero-shot translations. In this paper, we explore ways to improve them. We argue that multilingual NMT requires stronger modeling capacity to support language pairs with varying typological characteristics, and overcome this bottleneck via language-specific components and deepening NMT architectures. We identify the off-target translation issue (i.e. translating into a wrong target language) as the major source of the

inferior zero-shot performance, and propose random online backtranslation to enforce the translation of unseen training language pairs. Experiments on OPUS-100 (a novel multilingual dataset with 100 languages) show that our approach substantially narrows the performance gap with bilingual models in both one-to-many and many-to-many settings, and improves zero-shot performance by ~10 BLEU, approaching conventional pivot-based methods.

It's Easier to Translate out of English than into it: Measuring Neural Translation Difficulty by Cross-Mutual Information [Website][PDF]
Emanuele Bugliarello, Sabrina J. Mielke, Antonios Anastasopoulos, Ryan Cotterell, and Naoaki Okazaki
9:00–10:00

The performance of neural machine translation systems is commonly evaluated in terms of BLEU. However, due to its reliance on target language properties and generation, the BLEU metric does not allow an assessment of which translation directions are more difficult to model. In this paper, we propose cross-mutual information (XMI): an asymmetric information-theoretic metric of machine translation difficulty that exploits the probabilistic nature of most neural machine translation models. XMI allows us to better evaluate the difficulty of translating text into the target language while controlling for the difficulty of the target-side generation component independent of the translation task. We then present the first systematic and controlled study of cross-lingual translation difficulties using modern neural translation systems. Code for replicating our experiments is available online at https://github.com/e-bug/nmt-difficulty.

Language-aware Interlingua for Multilingual Neural Machine Translation [Website][PDF]
Changfeng Zhu, Heng Yu, Shanbo Cheng, and Weihua Luo 9:00–10:00

Multilingual neural machine translation (NMT) has led to impressive accuracy improvements in low-resource scenarios by sharing common linguistic information across languages. However, the traditional multilingual model fails to capture the diversity and specificity of different languages, resulting in inferior performance compared with individual models that are sufficiently trained. In this paper, we incorporate a language-aware interlingua into the Encoder-Decoder architecture. The interlingual network enables the model to learn a language-independent representation from the semantic spaces of different languages, while still allowing for language-specific specialization of a particular language-pair. Experiments show that our proposed method achieves remarkable improvements over state-of-the-art multilingual NMT baselines and produces comparable performance with strong individual models.

Norm-Based Curriculum Learning for Neural Machine Translation [Website][PDF]
Xuebo Liu, Houtim Lai, Derek F. Wong, and Lidia S. Chao 9:00–10:00

A neural machine translation (NMT) system is expensive to train, especially with high-resource settings. As the NMT architectures become deeper and wider, this issue gets worse and worse. In this paper, we aim to improve the efficiency of training an NMT by introducing a novel norm-based curriculum learning method. We use the norm (aka length or module) of a word embedding as a measure of 1) the difficulty of the sentence, 2) the competence of the model, and 3) the weight of the sentence. The norm-based sentence difficulty takes the advantages of both linguistically motivated and model-based sentence difficulties. It is easy to determine and contains learning-dependent features. The norm-based model competence makes NMT learn the curriculum in a fully automated way, while the norm-based sentence weight further enhances the learning of the vector representation of the NMT. Experimental results for the WMT'14 English-German and WMT'17 Chinese-English translation tasks demonstrate that the proposed method outperforms strong baselines in terms of BLEU score (+1.17/+1.56) and training speedup (2.22x/3.33x).

On the Limitations of Cross-lingual Encoders as Exposed by Reference-Free Machine Translation Evaluation [Website][PDF]
Wei Zhao, Goran Glavaš, Maxime Peyrard, Yang Gao, Robert West, and Steffen Eger 9:00–10:00

Evaluation of cross-lingual encoders is usually performed either via zero-shot cross-lingual transfer in supervised downstream tasks or via unsupervised cross-lingual textual similarity. In this paper, we concern ourselves with reference-free machine translation (MT) evaluation where we directly compare source texts to (sometimes low-quality) system translations, which represents a natural adversarial setup for multilingual encoders. Reference-free evaluation holds the promise of web-scale comparison of MT systems. We systematically investigate a range of metrics based on state-of-the-art cross-lingual semantic representations obtained with pretrained M-BERT and LASER. We find that they perform poorly as semantic encoders for reference-free MT evaluation and identify their two key limitations, namely, (a) a semantic mismatch between representations of mutual translations and, more prominently, (b) the inability to punish "translationese", i.e., low-quality literal translations. We propose two partial remedies: (1) post-hoc re-alignment of the vector spaces and (2) coupling of semantic-similarity based metrics with target-side language modeling. In segment-level MT evaluation, our best metric surpasses reference-based BLEU by 5.7 correlation points.

Parallel Sentence Mining by Constrained Decoding [Website][PDF]
Pinzhen Chen, Nikolay Bogoychev, Kenneth Heafield, and Faheem Kirefu 9:00–10:00

We present a novel method to extract parallel sentences from two monolingual corpora, using neural machine translation. Our method relies on translating sentences in one corpus, but constraining the decoding by a prefix tree built on the other corpus. We argue that a neural machine translation system by itself can be a sentence similarity scorer and it efficiently approximates pairwise comparison with a modified beam search. When benchmarked on the BUCC shared task, our method achieves results comparable to other submissions.

Self-Attention with Cross-Lingual Position Representation [Website][PDF]
Liang Ding, Longyue Wang, and Dacheng Tao 9:00–10:00

Position encoding (PE), an essential part of self-attention networks (SANs), is used to preserve the word order in-

formation for natural language processing tasks, generating fixed position indices for input sequences. However, in cross-lingual scenarios, e.g. machine translation, the PEs of source and target sentences are modeled independently. Due to word order divergences in different languages, modeling the cross-lingual positional relationships might help SANs tackle this problem. In this paper, we augment SANs with *cross-lingual position representations* to model the bilingually aware latent structure for the input sentence. Specifically, we utilize bracketing transduction grammar (BTG)-based reordering information to encourage SANs to learn bilingual diagonal alignments. Experimental results on WMT'14 English⇒German, WAT'17 Japanese⇒English, and WMT'17 Chinese⇔English translation tasks demonstrate that our approach significantly and consistently improves translation quality over strong baselines. Extensive analyses confirm that the performance gains come from the cross-lingual information.

"You Sound Just Like Your Father" Commercial Machine Translation Systems Include Stylistic Biases
[Website] [PDF]
Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari 9:00–10:00

The main goal of machine translation has been to convey the correct content. Stylistic considerations have been at best secondary. We show that as a consequence, the output of three commercial machine translation systems (Bing, DeepL, Google) make demographically diverse samples from five languages "sound" older and more male than the original. Our findings suggest that translation models reflect demographic bias in the training data. This opens up interesting new research avenues in machine translation to take stylistic considerations into account.

Session 2B: NLP Applications-2

Fast and Accurate Deep Bidirectional Language Representations for Unsupervised Learning [Website][PDF]
Joongbo Shin, Yoonhyung Lee, Seunghyun Yoon, and Kyomin Jung 9:00–10:00

Even though BERT has achieved successful performance improvements in various supervised learning tasks, BERT is still limited by repetitive inferences on unsupervised tasks for the computation of contextual language representations. To resolve this limitation, we propose a novel deep bidirectional language model called a Transformer-based Text Autoencoder (T-TA). The T-TA computes contextual language representations without repetition and displays the benefits of a deep bidirectional architecture, such as that of BERT. In computation time experiments in a CPU environment, the proposed T-TA performs over six times faster than the BERT-like model on a reranking task and twelve times faster on a semantic similarity task. Furthermore, the T-TA shows competitive or even better accuracies than those of BERT on the above tasks. Code is available at https://github.com/joongbo/tta.

Fine-grained Interest Matching for Neural News Recommendation [Website][PDF]
Heyuan Wang, Fangzhao Wu, Zheng Liu, and Xing Xie 9:00–10:00

Personalized news recommendation is a critical technology to improve users' online news reading experience. The core of news recommendation is accurate matching between user's interests and candidate news. The same user usually has diverse interests that are reflected in different news she has browsed. Meanwhile, important semantic features of news are implied in text segments of different granularities. Existing studies generally represent each user as a single vector and then match the candidate news vector, which may lose fine-grained information for recommendation. In this paper, we propose FIM, a Fine-grained Interest Matching method for neural news recommendation. Instead of aggregating user's all historical browsed news into a unified vector, we hierarchically construct multi-level representations for each news via stacked dilated convolutions. Then we perform fine-grained matching between segment pairs of each browsed news and the candidate news at each semantic level. High-order salient signals are then identified by resembling the hierarchy of image recognition for final click prediction. Extensive experiments on a real-world dataset from MSN news validate the effectiveness of our model on news recommendation.

Interpretable Operational Risk Classification with Semi-Supervised Variational Autoencoder [Website][PDF]
Fan Zhou, Shengming Zhang, and Yi Yang 9:00–10:00

Operational risk management is one of the biggest challenges nowadays faced by financial institutions. There are several major challenges of building a text classification system for automatic operational risk prediction, including imbalanced labeled/unlabeled data and lacking interpretability. To tackle these challenges, we present a semi-supervised text classification framework that integrates multi-head attention mechanism with Semi-supervised variational inference for Operational Risk Classification (SemiORC). We empirically evaluate the framework on a real-world dataset. The results demonstrate that our method can better utilize unlabeled data and learn visually interpretable document representations. SemiORC also outperforms other baseline methods on operational risk classification.

Interpreting Twitter User Geolocation [Website][PDF]
Ting Zhong, Tianliang Wang, Fan Zhou, Goce Trajcevski, Kunpeng Zhang, and Yi Yang 9:00–10:00

Identifying user geolocation in online social networks is an essential task in many location-based applications. Existing methods rely on the similarity of text and network structure, however, they suffer from a lack of interpretability on the corresponding results, which is crucial for understanding model behavior. In this work, we adopt influence functions to interpret the behavior of GNN-based models by identifying the importance of training users when predicting the locations of the testing users. This methodology helps with providing meaningful explanations on prediction results. Furthermore, it also initiates an attempt to uncover the so-called "black-box" GNN-based models by investigating the effect of individual nodes.

MMPE: A Multi-Modal Interface for Post-Editing Machine Translation [Website][PDF]
Nico Herbig, Tim Düwel, Santanu Pal, Kalliopi Meladaki, Mahsa Monshizadeh, Antonio Krüger, and Josef van Genabith 9:00–10:00

Current advances in machine translation (MT) increase the need for translators to switch from traditional translation to post-editing (PE) of machine-translated text, a process that saves time and reduces errors. This affects the design of translation interfaces, as the task changes from mainly generating text to correcting errors within otherwise helpful translation proposals. Since this paradigm shift offers potential for modalities other than mouse and keyboard, we present MMPE, the first prototype to combine traditional input modes with pen, touch, and speech modalities for PE of MT. The results of an evaluation with professional translators suggest that pen and touch interaction are suitable for deletion and reordering tasks, while they are of limited use for longer insertions. On the other hand, speech and multi-modal combinations of select & speech are considered suitable for replacements and insertions but offer less potential for deletion and reordering. Overall, participants were enthusiastic about the new modalities and saw them as good extensions to mouse & keyboard, but not as a complete substitute.

Modeling Code-Switch Languages Using Bilingual Parallel Corpus [Website][PDF]
Grandee Lee and Haizhou Li 9:00–10:00

Language modeling is the technique to estimate the probability of a sequence of words. A bilingual language model is expected to model the sequential dependency for words across languages, which is difficult due to the inherent lack of suitable training data as well as diverse syntactic structure across languages. We propose a bilingual attention

language model (BALM) that simultaneously performs language modeling objective with a quasi-translation objective to model both the monolingual as well as the cross-lingual sequential dependency. The attention mechanism learns the bilingual context from a parallel corpus. BALM achieves state-of-the-art performance on the SEAME code-switch database by reducing the perplexity of 20.5% over the best-reported result. We also apply BALM in bilingual lexicon induction, and language normalization tasks to validate the idea.

SpellGCN: Incorporating Phonological and Visual Similarities into Language Models for Chinese Spelling Check [Website] [PDF]
Xingyi Cheng, Weidi Xu, Kunlong Chen, Shaohua Jiang, Feng Wang, Taifeng Wang, Wei Chu, and Yuan Qi 9:00–10:00

Chinese Spelling Check (CSC) is a task to detect and correct spelling errors in Chinese natural language. Existing methods have made attempts to incorporate the similarity knowledge between Chinese characters. However, they take the similarity knowledge as either an external input resource or just heuristic rules. This paper proposes to incorporate phonological and visual similarity knowledge into language models for CSC via a specialized graph convolutional network (SpellGCN). The model builds a graph over the characters, and SpellGCN is learned to map this graph into a set of inter-dependent character classifiers. These classifiers are applied to the representations extracted by another network, such as BERT, enabling the whole network to be end-to-end trainable. Experiments are conducted on three human-annotated datasets. Our method achieves superior performance against previous models by a large margin.

Spelling Error Correction with Soft-Masked BERT [Website] [PDF]
Shaohua Zhang, Haoran Huang, Jicong Liu, and Hang Li 9:00–10:00

Spelling error correction is an important yet challenging task because a satisfactory solution of it essentially needs human-level language understanding ability. Without loss of generality we consider Chinese spelling error correction (CSC) in this paper. A state-of-the-art method for the task selects a character from a list of candidates for correction (including non-correction) at each position of the sentence on the basis of BERT, the language representation model. The accuracy of the method can be sub-optimal, however, because BERT does not have sufficient capability to detect whether there is an error at each position, apparently due to the way of pre-training it using mask language modeling. In this work, we propose a novel neural architecture to address the aforementioned issue, which consists of a network for error detection and a network for error correction based on BERT, with the former being connected to the latter with what we call soft-masking technique. Our method of using 'Soft-Masked BERT' is general, and it may be employed in other language detection-correction problems. Experimental results on two datasets, including one large dataset which we create and plan to release, demonstrate that the performance of our proposed method is significantly better than the baselines including the one solely based on BERT.

Session 2B: Resources and Evaluation-3

A Diverse Corpus for Evaluating and Developing English Math Word Problem Solvers [Website][PDF]
Shen-yun Miao, Chao-Chun Liang, and Keh-Yih Su 9:00–10:00

We present ASDiv (Academia Sinica Diverse MWP Dataset), a diverse (in terms of both language patterns and problem types) English math word problem (MWP) corpus for evaluating the capability of various MWP solvers. Existing MWP corpora for studying AI progress remain limited either in language usage patterns or in problem types. We thus present a new English MWP corpus with 2,305 MWPs that cover more text patterns and most problem types taught in elementary school. Each MWP is annotated with its problem type and grade level (for indicating the level of difficulty). Furthermore, we propose a metric to measure the lexicon usage diversity of a given MWP corpus, and demonstrate that ASDiv is more diverse than existing corpora. Experiments show that our proposed corpus reflects the true capability of MWP solvers more faithfully.

A Monolingual Approach to Contextualized Word Embeddings for Mid-Resource Languages [Website][PDF]
Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot 9:00–10:00

We use the multilingual OSCAR corpus, extracted from Common Crawl via language classification, filtering and cleaning, to train monolingual contextualized word embeddings (ELMo) for five mid-resource languages. We then compare the performance of OSCAR-based and Wikipedia-based ELMo embeddings for these languages on the part-of-speech tagging and parsing tasks. We show that, despite the noise in the Common-Crawl-based OSCAR data, embeddings trained on OSCAR perform much better than monolingual embeddings trained on Wikipedia. They actually equal or improve the current state of the art in tagging and parsing for all five languages. In particular, they also improve over multilingual Wikipedia-based contextual embeddings (multilingual BERT), which almost always constitutes the previous state of the art, thereby showing that the benefit of a larger, more diverse corpus surpasses the cross-lingual benefit of multilingual embedding architectures.

Improving Image Captioning Evaluation by Considering Inter References Variance [Website][PDF]
Yanzhi Yi, Hangyu Deng, and Jinglu Hu 9:00–10:00

Evaluating image captions is very challenging partially due to the fact that there are multiple correct captions for every single image. Most of the existing one-to-one metrics operate by penalizing mismatches between reference and generative caption without considering the intrinsic variance between ground truth captions. It usually leads to over-penalization and thus a bad correlation to human judgment. Recently, the latest one-to-one metric BERTScore can achieve high human correlation in system-level tasks while some issues can be fixed for better performance. In this paper, we propose a novel metric based on BERTScore that could handle such a challenge and extend BERTScore with a few new features appropriately for image captioning evaluation. The experimental results show that our metric achieves state-of-the-art human judgment correlation.

Revisiting the Context Window for Cross-lingual Word Embeddings [Website][PDF]
Ryokan Ri and Yoshimasa Tsuruoka 9:00–10:00

Existing approaches to mapping-based cross-lingual word embeddings are based on the assumption that the source and target embedding spaces are structurally similar. The structures of embedding spaces largely depend on the co-occurrence statistics of each word, which the choice of context window determines. Despite this obvious connection between the context window and mapping-based cross-lingual embeddings, their relationship has been underexplored in prior work. In this work, we provide a thorough evaluation, in various languages, domains, and tasks, of bilingual embeddings trained with different context windows. The highlight of our findings is that increasing the size of both the source and target window sizes improves the performance of bilingual lexicon induction, especially the performance on frequent nouns.

Will-They-Won't-They: A Very Large Dataset for Stance Detection on Twitter [Website][PDF]
Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier 9:00–10:00

We present a new challenging stance detection dataset, called Will-They-Won't-They (WT—WT), which contains 51,284 tweets in English, making it by far the largest available dataset of the type. All the annotations are carried out by experts; therefore, the dataset constitutes a high-quality and reliable benchmark for future research in stance detection. Our experiments with a wide range of recent state-of-the-art stance detection systems show that the dataset poses a strong challenge to existing models in this domain.

Session 2B: Student Research Workshop

SCAR: Sentence Compression using Autoencoders for Reconstruction [Website][PDF]
Chanakya Malireddy, Tirth Maniar, and Manish Shrivastava 9:00–10:00

Sentence compression is the task of shortening a sentence while retaining its meaning. Most methods proposed for this task rely on labeled or paired corpora (containing pairs of verbose and compressed sentences), which is often expensive to collect. To overcome this limitation, we present a novel unsupervised deep learning framework (SCAR) for deletion-based sentence compression. SCAR is primarily composed of two encoder-decoder pairs: a compressor and a reconstructor. The compressor masks the input, and the reconstructor tries to regenerate it. The model is entirely trained on unlabeled data and does not require additional inputs such as explicit syntactic information or optimal compression length. SCAR's merit lies in the novel Linkage Loss function, which correlates the compressor and its effect on reconstruction, guiding it to drop inferable tokens. SCAR achieves higher ROUGE scores on benchmark datasets than the existing state-of-the-art methods and baselines. We also conduct a user study to demonstrate the application of our model as a text highlighting system. Using our model to underscore salient information facilitates speed-reading and reduces the time required to skim a document.

Feature Difference Makes Sense: A medical image captioning model exploiting feature difference and tag information [Website][PDF]
Hyeryun Park, Kyungmo Kim, Jooyoung Yoon, Seongkeun Park, and Jinwook Choi 9:00–10:00

Medical image captioning can reduce the workload of physicians and save time and expense by automatically generating reports. However, current datasets are small and limited, creating additional challenges for researchers. In this study, we propose a feature difference and tag information combined long short-term memory (LSTM) model for chest x-ray report generation. A feature vector extracted from the image conveys visual information, but its ability to describe the image is limited. Other image captioning studies exhibited improved performance by exploiting feature differences, so the proposed model also utilizes them. First, we propose a difference and tag (DiTag) model containing the difference between the patient and normal images. Then, we propose a multi-difference and tag (mDiTag) model that also contains information about low-level differences, such as contrast, texture, and localized area. Evaluation of the proposed models demonstrates that the mDiTag model provides more information to generate captions and outperforms all other models.

Multi-Task Neural Model for Agglutinative Language Translation [Website][PDF]
Yirong Pan, Xiao Li, Yating Yang, and Rui Dong 9:00–10:00

Neural machine translation (NMT) has achieved impressive performance recently by using large-scale parallel corpora. However, it struggles in the low-resource and morphologically-rich scenarios of agglutinative language translation task. Inspired by the finding that monolingual data can greatly improve the NMT performance, we propose a multi-task neural model that jointly learns to perform bi-directional translation and agglutinative language stemming. Our approach employs the shared encoder and decoder to train a single model without changing the standard NMT architecture but instead adding a token before each source-side sentence to specify the desired target outputs of the two different tasks. Experimental results on Turkish-English and Uyghur-Chinese show that our proposed approach can significantly improve the translation performance on agglutinative languages by using a small amount of monolingual data.

Considering Likelihood in NLP Classification Explanations with Occlusion and Language Modeling [Website][PDF]
David Harbecke and Christoph Alt 9:00–10:00

Recently, state-of-the-art NLP models gained an increasing syntactic and semantic understanding of language, and explanation methods are crucial to understand their decisions. Occlusion is a well established method that provides explanations on discrete language data, e.g. by removing a language unit from an input and measuring the impact on a model's decision. We argue that current occlusion-based methods often produce invalid or syntactically incorrect language data, neglecting the improved abilities of recent NLP models. Furthermore, gradient-based explanation methods disregard the discrete distribution of data in NLP. Thus, we propose OLM: a novel explanation method that combines occlusion and language models to sample valid and syntactically correct replacements with high likelihood, given the context of the original input. We lay out a theoretical foundation that alleviates these weaknesses of other explanation methods in NLP and provide results that underline the importance of considering data likelihood in occlusion-based explanation.

Session 2B: Theory and Formalism in NLP (Linguistic and Mathematical)-2

A Three-Parameter Rank-Frequency Relation in Natural Languages [Website][PDF]
Chenchen Ding, Masao Utiyama, and Eiichiro Sumita 9:00–10:00

We present that, the rank-frequency relation in textual data follows $f \propto r^{-\alpha}(r+\gamma)^{-\beta}$, where f is the token frequency and r is the rank by frequency, with (α, β, γ) as parameters. The formulation is derived based on the empirical observation that $d^2(x+y)/dx^2$ is a typical impulse function, where $(x, y) = (\log r, \log f)$. The formulation is the power law when $\beta = 0$ and the Zipf–Mandelbrot law when $\alpha = 0$. We illustrate that α is related to the analytic features of syntax and $\beta+\gamma$ to those of morphology in natural languages from an investigation of multilingual corpora.

Dice Loss for Data-imbalanced NLP Tasks [Website][PDF]
Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li 9:00–10:00

Many NLP tasks such as tagging and machine reading comprehension are faced with the severe data imbalance issue: negative examples significantly outnumber positive examples, and the huge number of easy-negative examples overwhelms the training. The most commonly used cross entropy (CE) criteria is actually an accuracy-oriented objective, and thus creates a discrepancy between training and test: at training time, each training instance contributes equally to the objective function, while at test time F1 score concerns more about positive examples. In this paper, we propose to use dice loss in replacement of the standard cross-entropy objective for data-imbalanced NLP tasks. Dice loss is based on the Sørensen—Dice coefficient or Tversky index , which attaches similar importance to false positives and false negatives, and is more immune to the data-imbalance issue. To further alleviate the dominating influence from easy-negative examples in training, we propose to associate training examples with dynamically adjusted weights to deemphasize easy-negative examples. Theoretical analysis shows that this strategy narrows down the gap between the F1 score in evaluation and the dice loss in training. With the proposed training objective, we observe significant performance boost on a wide range of data imbalanced NLP tasks. Notably, we are able to achieve SOTA results on CTB5, CTB6 and UD1.4 for the part of speech tagging task; SOTA results on CoNLL03, OntoNotes5.0, MSRA and OntoNotes4.0 for the named entity recognition task; along with competitive results on the tasks of machine reading comprehension and paraphrase identification.

Language Models as an Alternative Evaluator of Word Order Hypotheses: A Case Study in Japanese [Website][PDF]
Tatsuki Kuribayashi, Takumi Ito, Jun Suzuki, and Kentaro Inui 9:00–10:00

We examine a methodology using neural language models (LMs) for analyzing the word order of language. This LM-based method has the potential to overcome the difficulties existing methods face, such as the propagation of preprocessor errors in count-based methods. In this study, we explore whether the LM-based method is valid for analyzing the word order. As a case study, this study focuses on Japanese due to its complex and flexible word order. To validate the LM-based method, we test (i) parallels between LMs and human word order preference, and (ii) consistency of the results obtained using the LM-based method with previous linguistic studies. Through our experiments, we tentatively conclude that LMs display sufficient word order knowledge for usage as an analysis tool. Finally, using the LM-based method, we demonstrate the relationship between the canonical word order and topicalization, which had yet to be analyzed by large-scale experiments.

[TACL] Theoretical Limitations of Self-Attention in Neural Sequence Models [Website][PDF]
Michael Hahn 9:00–10:00

Transformers are emerging as the new workhorse of NLP, showing great success across tasks. Unlike LSTMs, transformers process input sequences entirely through self-attention. Previous work has suggested that the computational capabilities of self-attention to process hierarchical structures are limited. In this work, we mathematically investigate the computational power of self-attention to model formal languages. Across both soft and hard attention, we show strong theoretical limitations of the computational abilities of self-attention, finding that it cannot model periodic finite-state languages, nor hierarchical structure, unless the number of layers or heads increases with input length. These limitations seem surprising given the practical success of self-attention and the prominent role assigned to hierarchical structure in linguistics, suggesting that natural language can be approximated well with models that are too weak for the formal languages typically assumed in theoretical linguistics.

Demo Session 2C

Time: *9:30–10:15*

pyBART: Evidence-based Syntactic Transformations for IE [Website][PDF]
Aryeh Tiktinsky, Yoav Goldberg, and Reut Tsarfaty

Syntactic dependencies can be predicted with high accuracy, and are useful for both machine-learned and pattern-based information extraction tasks. However, their utility can be improved. These syntactic dependencies are designed to accurately reflect syntactic relations, and they do not make semantic relations explicit. Therefore, these representations lack many explicit connections between content words, that would be useful for downstream applications. Proposals like English Enhanced UD improve the situation by extending universal dependency trees with additional explicit arcs. However, they are not available to Python users, and are also limited in coverage. We introduce a broad-coverage, data-driven and linguistically sound set of transformations, that makes event-structure and many lexical relations explicit. We present pyBART, an easy-to-use open-source Python library for converting English UD trees either to Enhanced UD graphs or to our representation. The library can work as a standalone package or be integrated within a spaCy NLP pipeline. When evaluated in a pattern-based relation extraction scenario, our representation results in higher extraction scores than Enhanced UD, while requiring fewer patterns.

Demo Session 3A

Time: *12:00–12:45*

EVIDENCEMINER: Textual Evidence Discovery for Life Sciences [Website][PDF]
Xuan Wang, Yingjun Guan, Weili Liu, Aabhas Chauhan, Enyi Jiang, Qi Li, David Liem, Dibakar Sigdel, John Caufield, Peipei Ping, and Jiawei Han

Traditional search engines for life sciences (e.g., PubMed) are designed for document retrieval and do not allow direct retrieval of specific statements. Some of these statements may serve as textual evidence that is key to tasks such as hypothesis generation and new finding validation. We present EVIDENCEMINER, a web-based system that lets users query a natural language statement and automatically retrieves textual evidence from a background corpora for life sciences. EVIDENCEMINER is constructed in a completely automated way without any human effort for training data annotation. It is supported by novel data-driven methods for distantly supervised named entity recognition and open information extraction. The entities and patterns are pre-computed and indexed offline to support fast online evidence retrieval. The annotation results are also highlighted in the original document for better visualization. EVIDENCEMINER also includes analytic functionalities such as the most frequent entity and relation summarization. EVIDENCEMINER can help scientists uncover important research issues, leading to more effective research and more in-depth quantitative analysis. The system of EVIDENCEMINER is available at https://evidenceminer.firebaseapp.com/.

Trialstreamer: Mapping and Browsing Medical Evidence in Real-Time [Website][PDF]
Benjamin Nye, Ani Nenkova, Iain Marshall, and Byron C. Wallace

We introduce Trialstreamer, a living database of clinical trial reports. Here we mainly describe the evidence extraction component; this extracts from biomedical abstracts key pieces of information that clinicians need when appraising the literature, and also the relations between these. Specifically, the system extracts descriptions of trial participants, the treatments compared in each arm (the interventions), and which outcomes were measured. The system then attempts to infer which interventions were reported to work best by determining their relationship with identified trial outcome measures. In addition to summarizing individual trials, these extracted data elements allow automatic synthesis of results across many trials on the same topic. We apply the system at scale to all reports of randomized controlled trials indexed in MEDLINE, powering the automatic generation of evidence maps, which provide a global view of the efficacy of different interventions combining data from all relevant clinical trials on a topic. We make all code and models freely available alongside a demonstration of the web interface.

Session 3A Overview – Monday, July 6, 2020 12:00–13:00

Track A *Cognitive Modeling and Psycholinguistics-3* Abstracts	A Systematic Assessment of Syntactic Generalization in Neural Language Models *Hu, Gauthier, Qian, Wilcox, and Levy* [Website][PDF]	Inflecting When There's No Majority: Limitations of Encoder-Decoder Neural Networks as Cognitive Models for German Plurals *McCurdy, Goldwater, and Lopez* [Website][PDF]	Overestimation of Syntactic Representation in Neural Language Models *Kodner and Gupta* [Website][PDF]	Suspense in Short Stories is Predicted By Uncertainty Reduction over Neural Story Representation *Wilmot and Keller* [Website][PDF]	You Don't Have Time to Read This: An Exploration of Document Reading Time Prediction *Weller, Hildebrandt, Reznik, Challis, Tass, Snell, and Seppi* [Website][PDF]
Track B *Computational Social Science and Social Media-3* Abstracts	Code-Switching Patterns Can Be an Effective Route to Improve Performance of Downstream NLP Applications: A Case Study of Humour, Sarcasm and Hate Speech Detection *Bansal, Garimella, Suhane, Patro, and Mukherjee* [Website][PDF]	DTCA: Decision Tree-based Co-Attention Networks for Explainable Claim Verification *Wu, Rao, Liang, and Nazir* [Website][PDF]	GCAN: Graph-aware Co-Attention Networks for Explainable Fake News Detection on Social Media *Lu and Li* [Website][PDF]		
Track C *Dialogue and Interactive Systems-5* Abstracts	A Generative Model for Joint Natural Language Understanding and Generation *Tseng, Cheng, Fang, and Vandyke* [Website][PDF]	Beyond User Self-Reported Likert Scale Ratings: A Comparison Model for Automatic Dialog Evaluation *Liang, Zou, and Yu* [Website][PDF]	Coach: A Coarse-to-Fine Approach for Cross-domain Slot Filling *Liu, Winata, Xu, and Fung* [Website][PDF]	Conversational Word Embedding for Retrieval-Based Dialog System *Ma, Cui, Liu, Wang, Wang, and Hu* [Website][PDF]	Few-shot Slot Tagging with Collapsed Dependency Transfer and Label-enhanced Task-adaptive Projection Network *Hou, Che, Lai, Zhou, Liu, Liu, and Liu* [Website][PDF]
	MuTual: A Dataset for Multi-Turn Dialogue Reasoning *Cui, Wu, Liu, Zhang, and Zhou* [Website][PDF]	PLATO: Pre-trained Dialogue Generation Model with Discrete Latent Variable *Bao, He, Wang, Wu, and Wang* [Website][PDF]	Paraphrase Augmented Task-Oriented Dialog Generation *Gao, Zhang, Ou, and Yu* [Website][PDF]	Span-ConveRT: Few-shot Span Extraction for Dialog with Pretrained Conversational Representations *Coope, Farghly, Gerz, Vulić, and Henderson* [Website][PDF]	You Impress Me: Dialogue Generation via Mutual Persona Perception *Liu, Chen, Chen, LOU, Chen, Zhou, and Zhang* [Website][PDF]
Track D *Generation-5* Abstracts	Automatic Detection of Generated Text is Easiest when Humans are Fooled *Ippolito, Duckworth, Callison-Burch, and Eck* [Website][PDF]	Fast and Accurate Non-Projective Dependency Tree Linearization *Yu, Tannert, Vu, and Kuhn* [Website][PDF]	Generating Diverse and Consistent QA pairs from Contexts with Information-Maximizing Hierarchical Conditional VAEs *Lee, Lee, Jeong, Kim, and Hwang* [Website][PDF]	Pre-train and Plug-in: Flexible Conditional Text Generation with Variational Auto-Encoders *Duan, Xu, Pei, Han, and Li* [Website][PDF]	Rigid Formats Controlled Text Generation *Li, Zhang, Liu, and Shi* [Website][PDF]

Track E *Information Retrieval and Text Mining-4* Abstracts	A Joint Model for Document Segmentation and Segment Labeling *Barrow, Jain, Morariu, Manjunatha, Oard, and Resnik* [Website][PDF]	An Online Semantic-enhanced Dirichlet Model for Short Text Stream Clustering *Kumar, Shao, Uddin, and Ali* [Website][PDF]	Dynamic Memory Induction Networks for Few-Shot Text Classification *Geng, Li, Li, Sun, and Zhu* [Website][PDF]	Exclusive Hierarchical Decoding for Deep Keyphrase Generation *Chen, Chan, Li, and King* [Website][PDF]	Generative Semantic Hashing Enhanced via Boltzmann Machines *Zheng, Su, Shen, and Chen* [Website][PDF]
	Hierarchy-Aware Global Model for Hierarchical Text Classification *Zhou, Ma, Long, Xu, Ding, Zhang, Xie, and Liu* [Website][PDF]	Keyphrase Generation for Scientific Document Retrieval *Boudin, Gallina, and Aizawa* [Website][PDF]	Unsupervised FAQ Retrieval with Question Generation and BERT *Mass, Carmeli, Roitman, and Konopnicki* [Website][PDF]		
Track F *Machine Translation-3* Abstracts	Boosting Neural Machine Translation with Similar Translations *XU, Crego, and Senellart* [Website][PDF]	End-to-End Neural Word Alignment Outperforms GIZA++ *Zenkel, Wuebker, and DeNero* [Website][PDF]	Enhancing Machine Translation with Dependency-Aware Self-Attention *Bugliarello and Okazaki* [Website][PDF]	Improving Massively Multilingual Neural Machine Translation and Zero-Shot Translation *Zhang, Williams, Titov, and Sennrich* [Website][PDF]	Jointly Masked Sequence-to-Sequence Model for Non-Autoregressive Neural Machine Translation *Guo, Xu, and Chen* [Website][PDF]
	Learning Source Phrase Representations for Neural Machine Translation *Xu, Genabith, Xiong, Liu, and Zhang* [Website][PDF]	Lipschitz Constrained Parameter Initialization for Deep Transformers *Xu, Liu, Genabith, Xiong, and Zhang* [Website][PDF]	Multi-Domain Neural Machine Translation with Word-Level Adaptive Layer-wise Domain Mixing *Jiang, Liang, Wang, and Zhao* [Website][PDF]	Self-Attention with Cross-Lingual Position Representation *Ding, Wang, and Tao* [Website][PDF]	"You Sound Just Like Your Father" Commercial Machine Translation Systems Include Stylistic Biases *Hovy, Bianchi, and Fornaciari* [Website][PDF]
Track G *Resources and Evaluation-4* Abstracts	A Monolingual Approach to Contextualized Word Embeddings for Mid-Resource Languages *Ortiz Suárez, Romary, and Sagot* [Website][PDF]	Building a User-Generated Content North-African Arabizi Treebank: Tackling Hell *Seddah, Essaidi, Fethi, Futeral, Muller, Ortiz Suárez, Sagot, and Srivastava* [Website][PDF]	Crawling and Preprocessing Mailing Lists At Scale for Dialog Analysis *Bevendorff, Al Khatib, Potthast, and Stein* [Website][PDF]	Fine-Grained Analysis of Cross-Linguistic Syntactic Divergences *Nikolaev, Arviv, Karidi, Kenneth, Mitnik, Saeboe, and Abend* [Website][PDF]	Generating Counter Narratives against Online Hate Speech: Data and Strategies *Tekiroğlu, Chung, and Guerini* [Website][PDF]
	KLEJ: Comprehensive Benchmark for Polish Language Understanding *Rybak, Mroczkowski, Tracz, and Gawlik* [Website][PDF]	Learning and Evaluating Emotion Lexicons for 91 Languages *Buechel, Rücker, and Hahn* [Website][PDF]	Multi-Hypothesis Machine Translation Evaluation *Fomicheva, Specia, and Guzmán* [Website][PDF]	Multimodal Quality Estimation for Machine Translation *Okabe, Blain, and Specia* [Website][PDF]	PuzzLing Machines: A Challenge on Learning From Small Data *Şahin, Kementchedjhieva, Rust, and Gurevych* [Website][PDF]

	The SOFC-Exp Corpus and Neural Approaches to Information Extraction in the Materials Science Domain *Friedrich, Adel, Tomazic, Hingerl, Benteau, Marusczyk, and Lange* [Website][PDF]	The TechQA Dataset *Castelli, Chakravarti, Dana, Ferritto, Florian, Franz, Garg, Khandelwal, McCarley, McCawley, Nasr, Pan, Pendus, Pitrelli, Pujar, Roukos, Sakrajda, Sil, Uceda-Sosa, Ward, and Zhang* [Website][PDF]	Will-They-Won't-They: A Very Large Dataset for Stance Detection on Twitter *Conforti, Berndt, Pilehvar, Giannitsarou, Toxvaerd, and Collier* [Website][PDF]	iSarcasm: A Dataset of Intended Sarcasm *Oprea and Magdy* [Website][PDF]	
Track H *Sentence Level-2* Abstracts	AMR Parsing via Graph-Sequence Iterative Inference *Cai and Lam* [Website][PDF]				
Track I *Student Research Workshop* Abstracts	Non-Topical Coherence in Social Talk: A Call for Dialogue Model Enrichment *Luu and Malamud* [Website][PDF]	Dominance as an Indicator of Rapport and Learning in Human-Agent Communication *Buddemeyer, Tian, and Walker* [Website]	SCAR: Sentence Compression using Autoencoders for Reconstruction *Malireddy, Maniar, and Shrivastava* [Website][PDF]	Why is penguin more similar to polar bear than to sea gull? Analyzing conceptual knowledge in distributional models *Sommerauer* [Website][PDF]	

Session 3A Details

Session 3A: Cognitive Modeling and Psycholinguistics-3

A Systematic Assessment of Syntactic Generalization in Neural Language Models [Website][PDF]
Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy 12:00–13:00

While state-of-the-art neural network models continue to achieve lower perplexity scores on language modeling benchmarks, it remains unknown whether optimizing for broad-coverage predictive performance leads to human-like syntactic knowledge. Furthermore, existing work has not provided a clear picture about the model properties required to produce proper syntactic generalizations. We present a systematic evaluation of the syntactic knowledge of neural language models, testing 20 combinations of model types and data sizes on a set of 34 English-language syntactic test suites. We find substantial differences in syntactic generalization performance by model architecture, with sequential models underperforming other architectures. Factorially manipulating model architecture and training dataset size (1M-40M words), we find that variability in syntactic generalization performance is substantially greater by architecture than by dataset size for the corpora tested in our experiments. Our results also reveal a dissociation between perplexity and syntactic generalization performance.

Inflecting When There's No Majority: Limitations of Encoder-Decoder Neural Networks as Cognitive Models for German Plurals [Website][PDF]
Kate McCurdy, Sharon Goldwater, and Adam Lopez 12:00–13:00

Can artificial neural networks learn to represent inflectional morphology and generalize to new words as human speakers do? Kirov and Cotterell (2018) argue that the answer is yes: modern Encoder-Decoder (ED) architectures learn human-like behavior when inflecting English verbs, such as extending the regular past tense form /-(e)d/ to novel words. However, their work does not address the criticism raised by Marcus et al. (1995): that neural models may learn to extend not the regular, but the most frequent class — and thus fail on tasks like German number inflection, where infrequent suffixes like /-s/ can still be productively generalized. To investigate this question, we first collect a new dataset from German speakers (production and ratings of plural forms for novel nouns) that is designed to avoid sources of information unavailable to the ED model. The speaker data show high variability, and two suffixes evince 'regular' behavior, appearing more often with phonologically atypical inputs. Encoder-decoder models do generalize the most frequently produced plural class, but do not show human-like variability or 'regular' extension of these other plural markers. We conclude that modern neural models may still struggle with minority-class generalization.

Overestimation of Syntactic Representation in Neural Language Models [Website][PDF]
Jordan Kodner and Nitish Gupta 12:00–13:00

With the advent of powerful neural language models over the last few years, research attention has increasingly focused on what aspects of language they represent that make them so successful. Several testing methodologies have been developed to probe models' syntactic representations. One popular method for determining a model's ability to induce syntactic structure trains a model on strings generated according to a template then tests the model's ability to distinguish such strings from superficially similar ones with different syntax. We illustrate a fundamental problem with this approach by reproducing positive results from a recent paper with two non-syntactic baseline language models: an n-gram model and an LSTM model trained on scrambled inputs.

Suspense in Short Stories is Predicted By Uncertainty Reduction over Neural Story Representation [Website][PDF]
David Wilmot and Frank Keller 12:00–13:00

Suspense is a crucial ingredient of narrative fiction, engaging readers and making stories compelling. While there is a vast theoretical literature on suspense, it is computationally not well understood. We compare two ways for modelling suspense: surprise, a backward-looking measure of how unexpected the current state is given the story so far; and uncertainty reduction, a forward-looking measure of how unexpected the continuation of the story is. Both can be computed either directly over story representations or over their probability distributions. We propose a hierarchical language model that encodes stories and computes surprise and uncertainty reduction. Evaluating against short stories annotated with human suspense judgements, we find that uncertainty reduction over representations is the best predictor, resulting in near human accuracy. We also show that uncertainty reduction can be used to predict suspenseful events in movie synopses.

You Don't Have Time to Read This: An Exploration of Document Reading Time Prediction [Website][PDF]
Orion Weller, Jordan Hildebrandt, Ilya Reznik, Christopher Challis, E. Shannon Tass, Quinn Snell, and Kevin Seppi 12:00–13:00

Predicting reading time has been a subject of much previous work, focusing on how different words affect human processing, measured by reading time. However, previous work has dealt with a limited number of participants as well as word level only predictions (i.e. predicting the time to read a single word). We seek to extend these works by examining whether or not document level predictions are effective, given additional information such as subject matter, font characteristics, and readability metrics. We perform a novel experiment to examine how different features of text contribute to the time it takes to read, distributing and collecting data from over a thousand participants. We then employ a large number of machine learning methods to predict a user's reading time. We find that despite

extensive research showing that word level reading time can be most effectively predicted by neural networks, larger scale text can be easily and most accurately predicted by one factor, the number of words.

Session 3A: Computational Social Science and Social Media-3

Code-Switching Patterns Can Be an Effective Route to Improve Performance of Downstream NLP Applications: A Case Study of Humour, Sarcasm and Hate Speech Detection [Website][PDF]
Srijan Bansal, Vishal Garimella, Ayush Suhane, Jasabanta Patro, and Animesh Mukherjee 12:00–13:00

In this paper, we demonstrate how code-switching patterns can be utilised to improve various downstream NLP applications. In particular, we encode various switching features to improve humour, sarcasm and hate speech detection tasks. We believe that this simple linguistic observation can also be potentially helpful in improving other similar NLP applications.

DTCA: Decision Tree-based Co-Attention Networks for Explainable Claim Verification [Website][PDF]
Lianwei Wu, Yuan Rao, yongqiang zhao yongqiang, Hao Liang, and Ambreen Nazir 12:00–13:00

Recently, many methods discover effective evidence from reliable sources by appropriate neural networks for explainable claim verification, which has been widely recognized. However, in these methods, the discovery process of evidence is nontransparent and unexplained. Simultaneously, the discovered evidence is aimed at the interpretability of the whole sequence of claims but insufficient to focus on the false parts of claims. In this paper, we propose a Decision Tree-based Co-Attention model (DTCA) to discover evidence for explainable claim verification. Specifically, we first construct Decision Tree-based Evidence model (DTE) to select comments with high credibility as evidence in a transparent and interpretable way. Then we design Co-attention Self-attention networks (CaSa) to make the selected evidence interact with claims, which is for 1) training DTE to determine the optimal decision thresholds and obtain more powerful evidence; and 2) utilizing the evidence to find the false parts in the claim. Experiments on two public datasets, RumourEval and PHEME, demonstrate that DTCA not only provides explanations for the results of claim verification but also achieves the state-of-the-art performance, boosting the F1-score by more than 3.11%, 2.41%, respectively.

GCAN: Graph-aware Co-Attention Networks for Explainable Fake News Detection on Social Media [Website][PDF]
Yi-Ju Lu and Cheng-Te Li 12:00–13:00

This paper solves the fake news detection problem under a more realistic scenario on social media. Given the source short-text tweet and the corresponding sequence of retweet users without text comments, we aim at predicting whether the source tweet is fake or not, and generating explanation by highlighting the evidences on suspicious retweeters and the words they concern. We develop a novel neural network-based model, Graph-aware Co-Attention Networks (GCAN), to achieve the goal. Extensive experiments conducted on real tweet datasets exhibit that GCAN can significantly outperform state-of-the-art methods by 16% in accuracy on average. In addition, the case studies also show that GCAN can produce reasonable explanations.

Session 3A: Dialogue and Interactive Systems-5

A Generative Model for Joint Natural Language Understanding and Generation [Website][PDF]
Bo-Hsiang Tseng, Jianpeng Cheng, Yimai Fang, and David Vandyke 12:00–13:00

Natural language understanding (NLU) and natural language generation (NLG) are two fundamental and related tasks in building task-oriented dialogue systems with opposite objectives: NLU tackles the transformation from natural language to formal representations, whereas NLG does the reverse. A key to success in either task is parallel training data which is expensive to obtain at a large scale. In this work, we propose a generative model which couples NLU and NLG through a shared latent variable. This approach allows us to explore both spaces of natural language and formal representations, and facilitates information sharing through the latent space to eventually benefit NLU and NLG. Our model achieves state-of-the-art performance on two dialogue datasets with both flat and tree-structured formal representations. We also show that the model can be trained in a semi-supervised fashion by utilising unlabelled data to boost its performance.

Beyond User Self-Reported Likert Scale Ratings: A Comparison Model for Automatic Dialog Evaluation [Website][PDF]
Weixin Liang, James Zou, and Zhou Yu 12:00–13:00

Open Domain dialog system evaluation is one of the most important challenges in dialog research. Existing automatic evaluation metrics, such as BLEU are mostly reference-based. They calculate the difference between the generated response and a limited number of available references. Likert-score based self-reported user rating is widely adopted by social conversational systems, such as Amazon Alexa Prize chatbots. However, self-reported user rating suffers from bias and variance among different users. To alleviate this problem, we formulate dialog evaluation as a comparison task. We also propose an automatic evaluation model CMADE (Comparison Model for Automatic Dialog Evaluation) that automatically cleans self-reported user ratings as it trains on them. Specifically, we first use a self-supervised method to learn better dialog feature representation, and then use KNN and Shapley to remove confusing samples. Our experiments show that CMADE achieves 89.2% accuracy in the dialog comparison task.

Coach: A Coarse-to-Fine Approach for Cross-domain Slot Filling [Website][PDF]
Zihan Liu, Genta Indra Winata, Peng Xu, and Pascale Fung 12:00–13:00

As an essential task in task-oriented dialog systems, slot filling requires extensive training data in a certain domain. However, such data are not always available. Hence, cross-domain slot filling has naturally arisen to cope with this data scarcity problem. In this paper, we propose a Coarse-to-fine approach (Coach) for cross-domain slot filling. Our model first learns the general pattern of slot entities by detecting whether the tokens are slot entities or not. It then predicts the specific types for the slot entities. In addition, we propose a template regularization approach to improve the adaptation robustness by regularizing the representation of utterances based on utterance templates. Experimental results show that our model significantly outperforms state-of-the-art approaches in slot filling. Furthermore, our model can also be applied to the cross-domain named entity recognition task, and it achieves better adaptation performance than other existing baselines. The code is available at https://github.com/zliucr/coach.

Conversational Word Embedding for Retrieval-Based Dialog System [Website][PDF]
Wentao Ma, Yiming Cui, Ting Liu, Dong Wang, Shijin Wang, and Guoping Hu 12:00–13:00

Human conversations contain many types of information, e.g., knowledge, common sense, and language habits. In this paper, we propose a conversational word embedding method named PR-Embedding, which utilizes the conversation pairs <post, reply> to learn word embedding. Different from previous works, PR-Embedding uses the vectors from two different semantic spaces to represent the words in post and reply.To catch the information among the pair, we first introduce the word alignment model from statistical machine translation to generate the cross-sentence window, then train the embedding on word-level and sentence-level.We evaluate the method on single-turn and multi-turn response selection tasks for retrieval-based dialog systems.The experiment results show that PR-Embedding can improve the quality of the selected response.

Few-shot Slot Tagging with Collapsed Dependency Transfer and Label-enhanced Task-adaptive Projection Network [Website][PDF]
Yutai Hou, Wanxiang Che, Yongkui Lai, Zhihan Zhou, Yijia Liu, Han Liu, and Ting Liu 12:00–13:00

In this paper, we explore the slot tagging with only a few labeled support sentences (a.k.a. few-shot). Few-shot slot tagging faces a unique challenge compared to the other fewshot classification problems as it calls for modeling the dependencies between labels. But it is hard to apply previously learned label dependencies to an unseen domain, due to the discrepancy of label sets. To tackle this, we introduce a collapsed dependency transfer mechanism into the conditional random field (CRF) to transfer abstract label dependency patterns as transition scores. In the few-shot setting, the emission score of CRF can be calculated as a word's similarity to the representation of each label. To calculate such similarity, we propose a Label-enhanced Task-Adaptive Projection Network (L-TapNet) based on the state-of-the-art few-shot classification model — TapNet, by leveraging label name semantics in representing labels. Experimental results show that our model significantly outperforms the strongest few-shot learning baseline by 14.64 F1 scores in the one-shot setting.

MuTual: A Dataset for Multi-Turn Dialogue Reasoning [Website][PDF]
Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou 12:00–13:00

Non-task oriented dialogue systems have achieved great success in recent years due to largely accessible conversation data and the development of deep learning techniques. Given a context, current systems are able to yield a relevant

and fluent response, but sometimes make logical mistakes because of weak reasoning capabilities. To facilitate the conversation reasoning research, we introduce MuTual, a novel dataset for Multi-Turn dialogue Reasoning, consisting of 8,860 manually annotated dialogues based on Chinese student English listening comprehension exams. Compared to previous benchmarks for non-task oriented dialogue systems, MuTual is much more challenging since it requires a model that be able to handle various reasoning problems. Empirical results show that state-of-the-art methods only reach 71%, which is far behind human performance of 94%, indicating that there is ample room for improving reasoning ability.

PLATO: Pre-trained Dialogue Generation Model with Discrete Latent Variable [Website][PDF]
Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang 12:00–13:00

Pre-training models have been proved effective for a wide range of natural language processing tasks. Inspired by this, we propose a novel dialogue generation pre-training framework to support various kinds of conversations, including chit-chat, knowledge grounded dialogues, and conversational question answering. In this framework, we adopt flexible attention mechanisms to fully leverage the bi-directional context and the uni-directional characteristic of language generation. We also introduce discrete latent variables to tackle the inherent one-to-many mapping problem in response generation. Two reciprocal tasks of response generation and latent act recognition are designed and carried out simultaneously within a shared network. Comprehensive experiments on three publicly available datasets verify the effectiveness and superiority of the proposed framework.

Paraphrase Augmented Task-Oriented Dialog Generation [Website][PDF]
Silin Gao, Yichi Zhang, Zhijian Ou, and Zhou Yu 12:00–13:00

Neural generative models have achieved promising performance on dialog generation tasks if given a huge data set. However, the lack of high-quality dialog data and the expensive data annotation process greatly limit their application in real world settings. We propose a paraphrase augmented response generation (PARG) framework that jointly trains a paraphrase model and a response generation model to improve the dialog generation performance. We also design a method to automatically construct paraphrase training data set based on dialog state and dialog act labels. PARG is applicable to various dialog generation models, such as TSCP (Lei et al., 2018) and DAMD (Zhang et al., 2019). Experimental results show that the proposed framework improves these state-of-the-art dialog models further on CamRest676 and MultiWOZ. PARG also outperforms other data augmentation methods significantly in dialog generation tasks, especially under low resource settings.

Span-ConveRT: Few-shot Span Extraction for Dialog with Pretrained Conversational Representations [Website][PDF]
Samuel Coope, Tyler Farghly, Daniela Gerz, Ivan Vulić, and Matthew Henderson 12:00–13:00

We introduce Span-ConveRT, a light-weight model for dialog slot-filling which frames the task as a turn-based span extraction task. This formulation allows for a simple integration of conversational knowledge coded in large pretrained conversational models such as ConveRT (Henderson et al., 2019). We show that leveraging such knowledge in Span-ConveRT is especially useful for few-shot learning scenarios: we report consistent gains over 1) a span extractor that trains representations from scratch in the target domain, and 2) a BERT-based span extractor. In order to inspire more work on span extraction for the slot-filling task, we also release RESTAURANTS-8K, a new challenging data set of 8,198 utterances, compiled from actual conversations in the restaurant booking domain.

You Impress Me: Dialogue Generation via Mutual Persona Perception [Website][PDF]
Qian Liu, Yihong Chen, Bei Chen, Jian-Guang LOU, Zixuan Chen, Bin Zhou, and Dongmei Zhang 12:00–13:00

Despite the continuing efforts to improve the engagingness and consistency of chit-chat dialogue systems, the majority of current work simply focus on mimicking human-like responses, leaving understudied the aspects of modeling understanding between interlocutors. The research in cognitive science, instead, suggests that understanding is an essential signal for a high-quality chit-chat conversation. Motivated by this, we propose P^2 Bot, a transmitter-receiver based framework with the aim of explicitly modeling understanding. Specifically, P^2 Bot incorporates mutual persona perception to enhance the quality of personalized dialogue generation. Experiments on a large public dataset, Persona-Chat, demonstrate the effectiveness of our approach, with a considerable boost over the state-of-the-art baselines across both automatic metrics and human evaluations.

Session 3A: Generation-5

Automatic Detection of Generated Text is Easiest when Humans are Fooled [Website][PDF]
Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck 12:00–13:00

Recent advancements in neural language modelling make it possible to rapidly generate vast amounts of human-sounding text. The capabilities of humans and automatic discriminators to detect machine-generated text have been a large source of research interest, but humans and machines rely on different cues to make their decisions. Here, we perform careful benchmarking and analysis of three popular sampling-based decoding strategies—top-_k_, nucleus sampling, and untruncated random sampling—and show that improvements in decoding methods have primarily optimized for fooling humans. This comes at the expense of introducing statistical abnormalities that make detection easy for automatic systems. We also show that though both human and automatic detector performance improve with longer excerpt length, even multi-sentence excerpts can fool expert human raters over 30% of the time. Our findings reveal the importance of using both human and automatic detectors to assess the humanness of text generation systems.

Fast and Accurate Non-Projective Dependency Tree Linearization [Website][PDF]
Xiang Yu, Simon Tannert, Ngoc Thang Vu, and Jonas Kuhn 12:00–13:00

We propose a graph-based method to tackle the dependency tree linearization task. We formulate the task as a Traveling Salesman Problem (TSP), and use a biaffine attention model to calculate the edge costs. We facilitate the decoding by solving the TSP for each subtree and combining the solution into a projective tree. We then design a transition system as post-processing, inspired by non-projective transition-based parsing, to obtain non-projective sentences. Our proposed method outperforms the state-of-the-art linearizer while being 10 times faster in training and decoding.

Generating Diverse and Consistent QA pairs from Contexts with Information-Maximizing Hierarchical Conditional VAEs [Website][PDF]
Dong Bok Lee, Seanie Lee, Woo Tae Jeong, Donghwan Kim, and Sung Ju Hwang 12:00–13:00

One of the most crucial challenges in question answering (QA) is the scarcity of labeled data, since it is costly to obtain question-answer (QA) pairs for a target text domain with human annotation. An alternative approach to tackle the problem is to use automatically generated QA pairs from either the problem context or from large amount of unstructured texts (e.g. Wikipedia). In this work, we propose a hierarchical conditional variational autoencoder (HCVAE) for generating QA pairs given unstructured texts as contexts, while maximizing the mutual information between generated QA pairs to ensure their consistency. We validate our Information Maximizing Hierarchical Conditional Variational AutoEncoder (Info-HCVAE) on several benchmark datasets by evaluating the performance of the QA model (BERT-base) using only the generated QA pairs (QA-based evaluation) or by using both the generated and human-labeled pairs (semi-supervised learning) for training, against state-of-the-art baseline models. The results show that our model obtains impressive performance gains over all baselines on both tasks, using only a fraction of data for training.

Pre-train and Plug-in: Flexible Conditional Text Generation with Variational Auto-Encoders [Website][PDF]
Yu Duan, Canwen Xu, Jiaxin Pei, Jialong Han, and Chenliang Li 12:00–13:00

Conditional Text Generation has drawn much attention as a topic of Natural Language Generation (NLG) which provides the possibility for humans to control the properties of generated contents. Current conditional generation models cannot handle emerging conditions due to their joint end-to-end learning fashion. When a new condition added, these techniques require full retraining. In this paper, we present a new framework named Pre-train and Plug-in Variational Auto-Encoder (PPVAE) towards flexible conditional text generation. PPVAE decouples the text generation module from the condition representation module to allow "one-to-many" conditional generation. When a fresh condition emerges, only a lightweight network needs to be trained and works as a plug-in for PPVAE, which is efficient and desirable for real-world applications. Extensive experiments demonstrate the superiority of PPVAE against the existing alternatives with better conditionality and diversity but less training effort.

Rigid Formats Controlled Text Generation [Website][PDF]
Piji Li, Haisong Zhang, Xiaojiang Liu, and Shuming Shi 12:00–13:00

Neural text generation has made tremendous progress in various tasks. One common characteristic of most of the tasks is that the texts are not restricted to some rigid formats when generating. However, we may confront some special text paradigms such as Lyrics (assume the music score is given), Sonnet, SongCi (classical Chinese poetry of the Song dynasty), etc. The typical characteristics of these texts are in three folds: (1) They must comply fully with the rigid predefined formats. (2) They must obey some rhyming schemes. (3) Although they are restricted to some formats, the sentence integrity must be guaranteed. To the best of our knowledge, text generation based on the predefined rigid formats has not been well investigated. Therefore, we propose a simple and elegant framework named SongNet to tackle this problem. The backbone of the framework is a Transformer-based auto-regressive language model. Sets of symbols are tailor-designed to improve the modeling performance especially on format, rhyme, and sentence integrity. We improve the attention mechanism to impel the model to capture some future information on the format. A pre-training and fine-tuning framework is designed to further improve the generation quality. Extensive experiments conducted on two collected corpora demonstrate that our proposed framework generates significantly better results in terms of both automatic metrics and the human evaluation.

Session 3A: Information Retrieval and Text Mining-4

A Joint Model for Document Segmentation and Segment Labeling [Website][PDF]
Joe Barrow, Rajiv Jain, Vlad Morariu, Varun Manjunatha, Douglas Oard, and Philip Resnik 12:00–13:00

Text segmentation aims to uncover latent structure by dividing text from a document into coherent sections. Where previous work on text segmentation considers the tasks of document segmentation and segment labeling separately, we show that the tasks contain complementary information and are best addressed jointly. We introduce Segment Pooling LSTM (S-LSTM), which is capable of jointly segmenting a document and labeling segments. In support of joint training, we develop a method for teaching the model to recover from errors by aligning the predicted and ground truth segments. We show that S-LSTM reduces segmentation error by 30% on average, while also improving segment labeling.

An Online Semantic-enhanced Dirichlet Model for Short Text Stream Clustering [Website][PDF]
Jay Kumar, Junming Shao, Salah Uddin, and Wazir Ali 12:00–13:00

Clustering short text streams is a challenging task due to its unique properties: infinite length, sparse data representation and cluster evolution. Existing approaches often exploit short text streams in a batch way. However, determine the optimal batch size is usually a difficult task since we have no priori knowledge when the topics evolve. In addition, traditional independent word representation in graphical model tends to cause "term ambiguity" problem in short text clustering. Therefore, in this paper, we propose an Online Semantic-enhanced Dirichlet Model for short sext stream clustering, called OSDM, which integrates the word-occurance semantic information (i.e., context) into a new graphical model and clusters each arriving short text automatically in an online way. Extensive results have demonstrated that OSDM has better performance compared to many state-of-the-art algorithms on both synthetic and real-world data sets.

Dynamic Memory Induction Networks for Few-Shot Text Classification [Website][PDF]
Ruiying Geng, Binhua Li, Yongbin Li, Jian Sun, and Xiaodan Zhu 12:00–13:00

This paper proposes Dynamic Memory Induction Networks (DMIN) for few-short text classification. The model develops a dynamic routing mechanism over static memory, enabling it to better adapt to unseen classes, a critical capability for few-short classification. The model also expands the induction process with supervised learning weights and query information to enhance the generalization ability of meta-learning. The proposed model brings forward the state-of-the-art performance significantly by 2–4% improvement on the miniRCV1 and ODIC datasets. Detailed analysis is further performed to show how the proposed network achieves the new performance.

Exclusive Hierarchical Decoding for Deep Keyphrase Generation [Website][PDF]
Wang Chen, Hou Pong Chan, Piji Li, and Irwin King 12:00–13:00

Keyphrase generation (KG) aims to summarize the main ideas of a document into a set of keyphrases. A new setting is recently introduced into this problem, in which, given a document, the model needs to predict a set of keyphrases and simultaneously determine the appropriate number of keyphrases to produce. Previous work in this setting employs a sequential decoding process to generate keyphrases. However, such a decoding method ignores the intrinsic hierarchical compositionality existing in the keyphrase set of a document. Moreover, previous work tends to generate duplicated keyphrases, which wastes time and computing resources. To overcome these limitations, we propose an exclusive hierarchical decoding framework that includes a hierarchical decoding process and either a soft or a hard exclusion mechanism. The hierarchical decoding process is to explicitly model the hierarchical compositionality of a keyphrase set. Both the soft and the hard exclusion mechanisms keep track of previously-predicted keyphrases within a window size to enhance the diversity of the generated keyphrases. Extensive experiments on multiple KG benchmark datasets demonstrate the effectiveness of our method to generate less duplicated and more accurate keyphrases.

Generative Semantic Hashing Enhanced via Boltzmann Machines [Website][PDF]
Lin Zheng, Qinliang Su, Dinghan Shen, and Changyou Chen 12:00–13:00

Generative semantic hashing is a promising technique for large-scale information retrieval thanks to its fast retrieval speed and small memory footprint. For the tractability of training, existing generative-hashing methods mostly assume a factorized form for the posterior distribution, enforcing independence among the bits of hash codes. From the perspectives of both model representation and code space size, independence is always not the best assumption. In this paper, to introduce correlations among the bits of hash codes, we propose to employ the distribution of Boltzmann machine as the variational posterior. To address the intractability issue of training, we first develop an approximate method to reparameterize the distribution of a Boltzmann machine by augmenting it as a hierarchical concatenation of a Gaussian-like distribution and a Bernoulli distribution. Based on that, an asymptotically-exact lower bound is further derived for the evidence lower bound (ELBO). With these novel techniques, the entire model can be optimized efficiently. Extensive experimental results demonstrate that by effectively modeling correlations among different bits within a hash code, our model can achieve significant performance gains.

Hierarchy-Aware Global Model for Hierarchical Text Classification [Website][PDF]
Jie Zhou, Chunping Ma, Dingkun Long, Guangwei Xu, Ning Ding, Haoyu Zhang, Pengjun Xie, and Gongshen Liu 12:00–13:00

Hierarchical text classification is an essential yet challenging subtask of multi-label text classification with a taxonomic hierarchy. Existing methods have difficulties in modeling the hierarchical label structure in a global view. Furthermore, they cannot make full use of the mutual interactions between the text feature space and the label space. In this paper, we formulate the hierarchy as a directed graph and introduce hierarchy-aware structure encoders for mod-

eling label dependencies. Based on the hierarchy encoder, we propose a novel end-to-end hierarchy-aware global model (HiAGM) with two variants. A multi-label attention variant (HiAGM-LA) learns hierarchy-aware label embeddings through the hierarchy encoder and conducts inductive fusion of label-aware text features. A text feature propagation model (HiAGM-TP) is proposed as the deductive variant that directly feeds text features into hierarchy encoders. Compared with previous works, both HiAGM-LA and HiAGM-TP achieve significant and consistent improvements on three benchmark datasets.

Keyphrase Generation for Scientific Document Retrieval [Website][PDF]
Florian Boudin, Ygor Gallina, and Akiko Aizawa 12:00–13:00

Sequence-to-sequence models have lead to significant progress in keyphrase generation, but it remains unknown whether they are reliable enough to be beneficial for document retrieval. This study provides empirical evidence that such models can significantly improve retrieval performance, and introduces a new extrinsic evaluation framework that allows for a better understanding of the limitations of keyphrase generation models. Using this framework, we point out and discuss the difficulties encountered with supplementing documents with -not present in text- keyphrases, and generalizing models across domains. Our code is available at https://github.com/boudinfl/ir-using-kg

Unsupervised FAQ Retrieval with Question Generation and BERT [Website][PDF]
Yosi Mass, Boaz Carmeli, Haggai Roitman, and David Konopnicki 12:00–13:00

We focus on the task of Frequently Asked Questions (FAQ) retrieval. A given user query can be matched against the questions and/or the answers in the FAQ. We present a fully unsupervised method that exploits the FAQ pairs to train two BERT models. The two models match user queries to FAQ answers and questions, respectively. We alleviate the missing labeled data of the latter by automatically generating high-quality question paraphrases. We show that our model is on par and even outperforms supervised models on existing datasets.

Session 3A: Machine Translation-3

Boosting Neural Machine Translation with Similar Translations [Website][PDF]
Jitao XU, Josep Crego, and Jean Senellart 12:00–13:00

This paper explores data augmentation methods for training Neural Machine Translation to make use of similar translations, in a comparable way a human translator employs fuzzy matches. In particular, we show how we can simply present the neural model with information of both source and target sides of the fuzzy matches, we also extend the similarity to include semantically related translations retrieved using sentence distributed representations. We show that translations based on fuzzy matching provide the model with "copy" information while translations based on embedding similarities tend to extend the translation "context". Results indicate that the effect from both similar sentences are adding up to further boost accuracy, combine naturally with model fine-tuning and are providing dynamic adaptation for unseen translation pairs. Tests on multiple data sets and domains show consistent accuracy improvements. To foster research around these techniques, we also release an Open-Source toolkit with efficient and flexible fuzzy-match implementation.

End-to-End Neural Word Alignment Outperforms GIZA++ [Website][PDF]
Thomas Zenkel, Joern Wuebker, and John DeNero 12:00–13:00

Word alignment was once a core unsupervised learning task in natural language processing because of its essential role in training statistical machine translation (MT) models. Although unnecessary for training neural MT models, word alignment still plays an important role in interactive applications of neural machine translation, such as annotation transfer and lexicon injection. While statistical MT methods have been replaced by neural approaches with superior performance, the twenty-year-old GIZA++ toolkit remains a key component of state-of-the-art word alignment systems. Prior work on neural word alignment has only been able to outperform GIZA++ by using its output during training. We present the first end-to-end neural word alignment method that consistently outperforms GIZA++ on three data sets. Our approach repurposes a Transformer model trained for supervised translation to also serve as an unsupervised word alignment model in a manner that is tightly integrated and does not affect translation quality.

Enhancing Machine Translation with Dependency-Aware Self-Attention [Website][PDF]
Emanuele Bugliarello and Naoaki Okazaki 12:00–13:00

Most neural machine translation models only rely on pairs of parallel sentences, assuming syntactic information is automatically learned by an attention mechanism. In this work, we investigate different approaches to incorporate syntactic knowledge in the Transformer model and also propose a novel, parameter-free, dependency-aware self-attention mechanism that improves its translation quality, especially for long sentences and in low-resource scenarios. We show the efficacy of each approach on WMT English-German and English-Turkish, and WAT English-Japanese translation tasks.

Improving Massively Multilingual Neural Machine Translation and Zero-Shot Translation [Website][PDF]
Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich 12:00–13:00

Massively multilingual models for neural machine translation (NMT) are theoretically attractive, but often underperform bilingual models and deliver poor zero-shot translations. In this paper, we explore ways to improve them. We argue that multilingual NMT requires stronger modeling capacity to support language pairs with varying typological characteristics, and overcome this bottleneck via language-specific components and deepening NMT architectures. We identify the off-target translation issue (i.e. translating into a wrong target language) as the major source of the inferior zero-shot performance, and propose random online backtranslation to enforce the translation of unseen training language pairs. Experiments on OPUS-100 (a novel multilingual dataset with 100 languages) show that our approach substantially narrows the performance gap with bilingual models in both one-to-many and many-to-many settings, and improves zero-shot performance by ~10 BLEU, approaching conventional pivot-based methods.

Jointly Masked Sequence-to-Sequence Model for Non-Autoregressive Neural Machine Translation [Website][PDF]
Junliang Guo, Linli Xu, and Enhong Chen 12:00–13:00

The masked language model has received remarkable attention due to its effectiveness on various natural language processing tasks. However, few works have adopted this technique in the sequence-to-sequence models. In this work, we introduce a jointly masked sequence-to-sequence model and explore its application on non-autoregressive neural machine translation~(NAT). Specifically, we first empirically study the functionalities of the encoder and the decoder in NAT models, and find that the encoder takes a more important role than the decoder regarding the translation quality. Therefore, we propose to train the encoder more rigorously by masking the encoder input while training. As for the decoder, we propose to train it based on the consecutive masking of the decoder input with an n-gram loss function to alleviate the problem of translating duplicate words. The two types of masks are applied to the model jointly at the training stage. We conduct experiments on five benchmark machine translation tasks, and our model can achieve 27.69/32.24 BLEU scores on WMT14 English-German/German-English tasks with $5+$ times speed up compared with an autoregressive model.

Learning Source Phrase Representations for Neural Machine Translation [Website][PDF]
Hongfei Xu, Josef van Genabith, Deyi Xiong, Qiuhui Liu, and Jingyi Zhang 12:00–13:00

The Transformer translation model (Vaswani et al., 2017) based on a multi-head attention mechanism can be computed effectively in parallel and has significantly pushed forward the performance of Neural Machine Translation

(NMT). Though intuitively the attentional network can connect distant words via shorter network paths than RNNs, empirical analysis demonstrates that it still has difficulty in fully capturing long-distance dependencies (Tang et al., 2018). Considering that modeling phrases instead of words has significantly improved the Statistical Machine Translation (SMT) approach through the use of larger translation blocks ("phrases") and its reordering ability, modeling NMT at phrase level is an intuitive proposal to help the model capture long-distance relationships. In this paper, we first propose an attentive phrase representation generation mechanism which is able to generate phrase representations from corresponding token representations. In addition, we incorporate the generated phrase representations into the Transformer translation model to enhance its ability to capture long-distance relationships. In our experiments, we obtain significant improvements on the WMT 14 English-German and English-French tasks on top of the strong Transformer baseline, which shows the effectiveness of our approach. Our approach helps Transformer Base models perform at the level of Transformer Big models, and even significantly better for long sentences, but with substantially fewer parameters and training steps. The fact that phrase representations help even in the big setting further supports our conjecture that they make a valuable contribution to long-distance relations.

Lipschitz Constrained Parameter Initialization for Deep Transformers [Website][PDF]
Hongfei Xu, Qiuhui Liu, Josef van Genabith, Deyi Xiong, and Jingyi Zhang 12:00–13:00

The Transformer translation model employs residual connection and layer normalization to ease the optimization difficulties caused by its multi-layer encoder/decoder structure. Previous research shows that even with residual connection and layer normalization, deep Transformers still have difficulty in training, and particularly Transformer models with more than 12 encoder/decoder layers fail to converge. In this paper, we first empirically demonstrate that a simple modification made in the official implementation, which changes the computation order of residual connection and layer normalization, can significantly ease the optimization of deep Transformers. We then compare the subtle differences in computation order in considerable detail, and present a parameter initialization method that leverages the Lipschitz constraint on the initialization of Transformer parameters that effectively ensures training convergence. In contrast to findings in previous research we further demonstrate that with Lipschitz parameter initialization, deep Transformers with the original computation order can converge, and obtain significant BLEU improvements with up to 24 layers. In contrast to previous research which focuses on deep encoders, our approach additionally enables Transformers to also benefit from deep decoders.

Multi-Domain Neural Machine Translation with Word-Level Adaptive Layer-wise Domain Mixing [Website][PDF]
Haoming Jiang, Chen Liang, Chong Wang, and Tuo Zhao 12:00–13:00

Many multi-domain neural machine translation (NMT) models achieve knowledge transfer by enforcing one encoder to learn shared embedding across domains. However, this design lacks adaptation to individual domains. To overcome this limitation, we propose a novel multi-domain NMT model using individual modules for each domain, on which we apply word-level, adaptive and layer-wise domain mixing. We first observe that words in a sentence are often related to multiple domains. Hence, we assume each word has a domain proportion, which indicates its domain preference. Then word representations are obtained by mixing their embedding in individual domains based on their domain proportions. We show this can be achieved by carefully designing multi-head dot-product attention modules for different domains, and eventually taking weighted averages of their parameters by word-level layer-wise domain proportions. Through this, we can achieve effective domain knowledge sharing and capture fine-grained domain-specific knowledge as well. Our experiments show that our proposed model outperforms existing ones in several NMT tasks.

Self-Attention with Cross-Lingual Position Representation [Website][PDF]
Liang Ding, Longyue Wang, and Dacheng Tao 12:00–13:00

Position encoding (PE), an essential part of self-attention networks (SANs), is used to preserve the word order information for natural language processing tasks, generating fixed position indices for input sequences. However, in cross-lingual scenarios, e.g. machine translation, the PEs of source and target sentences are modeled independently. Due to word order divergences in different languages, modeling the cross-lingual positional relationships might help SANs tackle this problem. In this paper, we augment SANs with *cross-lingual position representations* to model the bilingually aware latent structure for the input sentence. Specifically, we utilize bracketing transduction grammar (BTG)-based reordering information to encourage SANs to learn bilingual diagonal alignments. Experimental results on WMT'14 English⇒German, WAT'17 Japanese⇒English, and WMT'17 Chinese⇔English translation tasks demonstrate that our approach significantly and consistently improves translation quality over strong baselines. Extensive analyses confirm that the performance gains come from the cross-lingual information.

"You Sound Just Like Your Father" Commercial Machine Translation Systems Include Stylistic Biases [Website][PDF]
Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari 12:00–13:00

The main goal of machine translation has been to convey the correct content. Stylistic considerations have been at best secondary. We show that as a consequence, the output of three commercial machine translation systems (Bing, DeepL, Google) make demographically diverse samples from five languages "sound" older and more male than the original. Our findings suggest that translation models reflect demographic bias in the training data. This opens up interesting new research avenues in machine translation to take stylistic considerations into account.

Session 3A: Resources and Evaluation-4

A Monolingual Approach to Contextualized Word Embeddings for Mid-Resource Languages [Website][PDF]
Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot 12:00–13:00

We use the multilingual OSCAR corpus, extracted from Common Crawl via language classification, filtering and cleaning, to train monolingual contextualized word embeddings (ELMo) for five mid-resource languages. We then compare the performance of OSCAR-based and Wikipedia-based ELMo embeddings for these languages on the part-of-speech tagging and parsing tasks. We show that, despite the noise in the Common-Crawl-based OSCAR data, embeddings trained on OSCAR perform much better than monolingual embeddings trained on Wikipedia. They actually equal or improve the current state of the art in tagging and parsing for all five languages. In particular, they also improve over multilingual Wikipedia-based contextual embeddings (multilingual BERT), which almost always constitutes the previous state of the art, thereby showing that the benefit of a larger, more diverse corpus surpasses the cross-lingual benefit of multilingual embedding architectures.

Building a User-Generated Content North-African Arabizi Treebank: Tackling Hell [Website][PDF]
Djamé Seddah, Farah Essaidi, Amal Fethi, Matthieu Futeral, Benjamin Muller, Pedro Javier Ortiz Suárez, Benoît Sagot, and Abhishek Srivastava 12:00–13:00

We introduce the first treebank for a romanized user-generated content variety of Algerian, a North-African Arabic dialect known for its frequent usage of code-switching. Made of 1500 sentences, fully annotated in morpho-syntax and Universal Dependency syntax, with full translation at both the word and the sentence levels, this treebank is made freely available. It is supplemented with 50k unlabeled sentences collected from Common Crawl and web-crawled data using intensive data-mining techniques. Preliminary experiments demonstrate its usefulness for POS tagging and dependency parsing. We believe that what we present in this paper is useful beyond the low-resource language community. This is the first time that enough unlabeled and annotated data is provided for an emerging user-generated content dialectal language with rich morphology and code switching, making it an challenging testbed for most recent NLP approaches.

Crawling and Preprocessing Mailing Lists At Scale for Dialog Analysis [Website][PDF]
Janek Bevendorff, Khalid Al Khatib, Martin Potthast, and Benno Stein 12:00–13:00

This paper introduces the Webis Gmane Email Corpus 2019, the largest publicly available and fully preprocessed email corpus to date. We crawled more than 153 million emails from 14,699 mailing lists and segmented them into semantically consistent components using a new neural segmentation model. With 96% accuracy on 15 classes of email segments, our model achieves state-of-the-art performance while being more efficient to train than previous ones. All data, code, and trained models are made freely available alongside the paper.

Fine-Grained Analysis of Cross-Linguistic Syntactic Divergences [Website][PDF]
Dmitry Nikolaev, Ofir Arviv, Taelin Karidi, Neta Kenneth, Veronika Mitnik, Lilja Maria Saeboe, and Omri Abend 12:00–13:00

The patterns in which the syntax of different languages converges and diverges are often used to inform work on cross-lingual transfer. Nevertheless, little empirical work has been done on quantifying the prevalence of different syntactic divergences across language pairs. We propose a framework for extracting divergence patterns for any language pair from a parallel corpus, building on Universal Dependencies. We show that our framework provides a detailed picture of cross-language divergences, generalizes previous approaches, and lends itself to full automation. We further present a novel dataset, a manually word-aligned subset of the Parallel UD corpus in five languages, and use it to perform a detailed corpus study. We demonstrate the usefulness of the resulting analysis by showing that it can help account for performance patterns of a cross-lingual parser.

Generating Counter Narratives against Online Hate Speech: Data and Strategies [Website][PDF]
Serra Sinem Tekiroğlu, Yi-Ling Chung, and Marco Guerini 12:00–13:00

Recently research has started focusing on avoiding undesired effects that come with content moderation, such as censorship and overblocking, when dealing with hatred online. The core idea is to directly intervene in the discussion with textual responses that are meant to counter the hate content and prevent it from further spreading. Accordingly, automation strategies, such as natural language generation, are beginning to be investigated. Still, they suffer from the lack of sufficient amount of quality data and tend to produce generic/repetitive responses. Being aware of the aforementioned limitations, we present a study on how to collect responses to hate effectively, employing large scale unsupervised language models such as GPT-2 for the generation of silver data, and the best annotation strategies/neural architectures that can be used for data filtering before expert validation/post-editing.

KLEJ: Comprehensive Benchmark for Polish Language Understanding [Website][PDF]
Piotr Rybak, Robert Mroczkowski, Janusz Tracz, and Ireneusz Gawlik 12:00–13:00

In recent years, a series of Transformer-based models unlocked major improvements in general natural language understanding (NLU) tasks. Such a fast pace of research would not be possible without general NLU benchmarks, which allow for a fair comparison of the proposed methods. However, such benchmarks are available only for a handful of languages. To alleviate this issue, we introduce a comprehensive multi-task benchmark for the Polish language understanding, accompanied by an online leaderboard. It consists of a diverse set of tasks, adopted from existing datasets for named entity recognition, question-answering, textual entailment, and others. We also introduce a new sentiment analysis task for the e-commerce domain, named Allegro Reviews (AR). To ensure a common evaluation

scheme and promote models that generalize to different NLU tasks, the benchmark includes datasets from varying domains and applications. Additionally, we release HerBERT, a Transformer-based model trained specifically for the Polish language, which has the best average performance and obtains the best results for three out of nine tasks. Finally, we provide an extensive evaluation, including several standard baselines and recently proposed, multilingual Transformer-based models.

Learning and Evaluating Emotion Lexicons for 91 Languages [Website][PDF]
Sven Buechel, Susanna Rücker, and Udo Hahn 12:00–13:00

Emotion lexicons describe the affective meaning of words and thus constitute a centerpiece for advanced sentiment and emotion analysis. Yet, manually curated lexicons are only available for a handful of languages, leaving most languages of the world without such a precious resource for downstream applications. Even worse, their coverage is often limited both in terms of the lexical units they contain and the emotional variables they feature. In order to break this bottleneck, we here introduce a methodology for creating almost arbitrarily large emotion lexicons for any target language. Our approach requires nothing but a source language emotion lexicon, a bilingual word translation model, and a target language embedding model. Fulfilling these requirements for 91 languages, we are able to generate representationally rich high-coverage lexicons comprising eight emotional variables with more than 100k lexical entries each. We evaluated the automatically generated lexicons against human judgment from 26 datasets, spanning 12 typologically diverse languages, and found that our approach produces results in line with state-of-the-art monolingual approaches to lexicon creation and even surpasses human reliability for some languages and variables. Code and data are available at https://github.com/JULIELab/MEmoLon archived under DOI 10.5281/zenodo.3779901.

Multi-Hypothesis Machine Translation Evaluation [Website][PDF]
Marina Fomicheva, Lucia Specia, and Francisco Guzmán 12:00–13:00

Reliably evaluating Machine Translation (MT) through automated metrics is a long-standing problem. One of the main challenges is the fact that multiple outputs can be equally valid. Attempts to minimise this issue include metrics that relax the matching of MT output and reference strings, and the use of multiple references. The latter has been shown to significantly improve the performance of evaluation metrics. However, collecting multiple references is expensive and in practice a single reference is generally used. In this paper, we propose an alternative approach: instead of modelling linguistic variation in human reference we exploit the MT model uncertainty to generate multiple diverse translations and use these: (i) as surrogates to reference translations; (ii) to obtain a quantification of translation variability to either complement existing metric scores or (iii) replace references altogether. We show that for a number of popular evaluation metrics our variability estimates lead to substantial improvements in correlation with human judgements of quality by up 15%.

Multimodal Quality Estimation for Machine Translation [Website][PDF]
Shu Okabe, Frédéric Blain, and Lucia Specia 12:00–13:00

We propose approaches to Quality Estimation (QE) for Machine Translation that explore both text and visual modalities for Multimodal QE. We compare various multimodality integration and fusion strategies. For both sentence-level and document-level predictions, we show that state-of-the-art neural and feature-based QE frameworks obtain better results when using the additional modality.

PuzzLing Machines: A Challenge on Learning From Small Data [Website][PDF]
Gözde Gül Şahin, Yova Kementchedjhieva, Phillip Rust, and Iryna Gurevych 12:00–13:00

Deep neural models have repeatedly proved excellent at memorizing surface patterns from large datasets for various ML and NLP benchmarks. They struggle to achieve human-like thinking, however, because they lack the skill of iterative reasoning upon knowledge. To expose this problem in a new light, we introduce a challenge on learning from small data, PuzzLing Machines, which consists of Rosetta Stone puzzles from Linguistic Olympiads for high school students. These puzzles are carefully designed to contain only the minimal amount of parallel text necessary to deduce the form of unseen expressions. Solving them does not require external information (e.g., knowledge bases, visual signals) or linguistic expertise, but meta-linguistic awareness and deductive skills. Our challenge contains around 100 puzzles covering a wide range of linguistic phenomena from 81 languages. We show that both simple statistical algorithms and state-of-the-art deep neural models perform inadequately on this challenge, as expected. We hope that this benchmark, available at https://ukplab.github.io/PuzzLing-Machines/, inspires further efforts towards a new paradigm in NLP—one that is grounded in human-like reasoning and understanding.

The SOFC-Exp Corpus and Neural Approaches to Information Extraction in the Materials Science Domain [Website][PDF]
Annemarie Friedrich, Heike Adel, Federico Tomazic, Johannes Hingerl, Renou Benteau, Anika Marusczyk, and Lukas Lange 12:00–13:00

This paper presents a new challenging information extraction task in the domain of materials science. We develop an annotation scheme for marking information on experiments related to solid oxide fuel cells in scientific publications, such as involved materials and measurement conditions. With this paper, we publish our annotation guidelines, as well as our SOFC-Exp corpus consisting of 45 open-access scholarly articles annotated by domain experts. A corpus and an inter-annotator agreement study demonstrate the complexity of the suggested named entity recognition and slot filling tasks as well as high annotation quality. We also present strong neural-network based models for a variety of tasks that can be addressed on the basis of our new data set. On all tasks, using BERT embeddings leads to large performance gains, but with increasing task complexity, adding a recurrent neural network on top seems beneficial. Our models will serve as competitive baselines in future work, and analysis of their performance highlights difficult cases when modeling the data and suggests promising research directions.

The TechQA Dataset [Website][PDF]
Vittorio Castelli, Rishav Chakravarti, Saswati Dana, Anthony Ferritto, Radu Florian, Martin Franz, Dinesh Garg, Dinesh Khandelwal, Scott McCarley, Michael McCawley, Mohamed Nasr, Lin Pan, Cezar Pendus, John Pitrelli, Saurabh Pujar, Salim Roukos, Andrzej Sakrajda, Avi Sil, Rosario Uceda-Sosa, Todd Ward, and Rong Zhang 12:00–13:00

We introduce TECHQA, a domain-adaptation question answering dataset for the technical support domain. The TECHQA corpus highlights two real-world issues from the automated customer support domain. First, it contains actual questions posed by users on a technical forum, rather than questions generated specifically for a competition or a task. Second, it has a real-world size — 600 training, 310 dev, and 490 evaluation question/answer pairs — thus reflecting the cost of creating large labeled datasets with actual data. Hence, TECHQA is meant to stimulate research in domain adaptation rather than as a resource to build QA systems from scratch. TECHQA was obtained by crawling the IBMDeveloper and DeveloperWorks forums for questions with accepted answers provided in an IBM Technote—a technical document that addresses a specific technical issue. We also release a collection of the 801,998 Technotes available on the web as of April 4, 2019 as a companion resource that can be used to learn representations of the IT domain language.

Will-They-Won't-They: A Very Large Dataset for Stance Detection on Twitter [Website][PDF]
Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier 12:00–13:00

We present a new challenging stance detection dataset, called Will-They-Won't-They (WT—WT), which contains 51,284 tweets in English, making it by far the largest available dataset of the type. All the annotations are carried out by experts; therefore, the dataset constitutes a high-quality and reliable benchmark for future research in stance detection. Our experiments with a wide range of recent state-of-the-art stance detection systems show that the dataset poses a strong challenge to existing models in this domain.

iSarcasm: A Dataset of Intended Sarcasm [Website][PDF]
Silviu Oprea and Walid Magdy 12:00–13:00

We consider the distinction between intended and perceived sarcasm in the context of textual sarcasm detection. The former occurs when an utterance is sarcastic from the perspective of its author, while the latter occurs when the utterance is interpreted as sarcastic by the audience. We show the limitations of previous labelling methods in capturing intended sarcasm and introduce the iSarcasm dataset of tweets labeled for sarcasm directly by their authors. Examining the state-of-the-art sarcasm detection models on our dataset showed low performance compared to previously studied datasets, which indicates that these datasets might be biased or obvious and sarcasm could be a phenomenon under-studied computationally thus far. By providing the iSarcasm dataset, we aim to encourage future NLP research to develop methods for detecting sarcasm in text as intended by the authors of the text, not as labeled under assumptions that we demonstrate to be sub-optimal.

Session 3A Semantics: Sentence Level-2

AMR Parsing via Graph-Sequence Iterative Inference [Website][PDF]
Deng Cai and Wai Lam 12:00–13:00

We propose a new end-to-end model that treats AMR parsing as a series of dual decisions on the input sequence and the incrementally constructed graph. At each time step, our model performs multiple rounds of attention, reasoning, and composition that aim to answer two critical questions: (1) which part of the input *sequence* to abstract; and (2) where in the output *graph* to construct the new concept. We show that the answers to these two questions are mutually causalities. We design a model based on iterative inference that helps achieve better answers in both perspectives, leading to greatly improved parsing accuracy. Our experimental results significantly outperform all previously reported SMATCH scores by large margins. Remarkably, without the help of any large-scale pre-trained language model (e.g., BERT), our model already surpasses previous state-of-the-art using BERT. With the help of BERT, we can push the state-of-the-art results to 80.2% on LDC2017T10 (AMR 2.0) and 75.4% on LDC2014T12 (AMR 1.0).

Session 3A: Student Research Workshop

Non-Topical Coherence in Social Talk: A Call for Dialogue Model Enrichment [Website] [PDF]
Alex Luu and Sophia A. Malamud 12:00–13:00

Current models of dialogue mainly focus on utterances within a topically coherent discourse segment, rather than new-topic utterances (NTUs), which begin a new topic not correlating with the content of prior discourse. As a result, these models may sufficiently account for discourse context of task-oriented but not social conversations. We conduct a pilot annotation study of NTUs as a first step towards a model capable of rationalizing conversational coherence in social talk. We start with the naturally occurring social dialogues in the Disco-SPICE corpus, annotated with discourse relations in the Penn Discourse Treebank and Cognitive approach to Coherence Relations frameworks. We first annotate content-based coherence relations that are not available in Disco-SPICE, and then heuristically identify NTUs, which lack a coherence relation to prior discourse. Based on the interaction between NTUs and their discourse context, we construct a classification for NTUs that actually convey certain non-topical coherence in social talk. This classification introduces new sequence-based social intents that traditional taxonomies of speech acts do not capture. The new findings advocates the development of a Bayesian game-theoretic model for social talk.

Dominance as an Indicator of Rapport and Learning in Human-Agent Communication [Website]
Amanda Buddemeyer, Xiaoyi Tian, and Erin Walker 12:00–13:00

Power dynamics in human-human communication can impact rapport-building and learning gains, but little is known about how power impacts human-agent communication. In this paper, we examine dominance behavior in utterances between middle-school students and a teachable robot as they work through math problems, as coded by Rogers and Farace's Relational Communication Control Coding Scheme (RCCCS). We hypothesize that relatively dominant students will show increased learning gains, as will students with greater dominance agreement with the robot. We also hypothesize that gender could be an indicator of differences in dominance behavior. We present a preliminary analysis of dominance characteristics in some of the transactions between robot and student. Ultimately, we hope to determine if manipulating the dominance behavior of a learning robot could support learning.

SCAR: Sentence Compression using Autoencoders for Reconstruction [Website] [PDF]
Chanakya Malireddy, Tirth Maniar, and Manish Shrivastava 12:00–13:00

Sentence compression is the task of shortening a sentence while retaining its meaning. Most methods proposed for this task rely on labeled or paired corpora (containing pairs of verbose and compressed sentences), which is often expensive to collect. To overcome this limitation, we present a novel unsupervised deep learning framework (SCAR) for deletion-based sentence compression. SCAR is primarily composed of two encoder-decoder pairs: a compressor and a reconstructor. The compressor masks the input, and the reconstructor tries to regenerate it. The model is entirely trained on unlabeled data and does not require additional inputs such as explicit syntactic information or optimal compression length. SCAR's merit lies in the novel Linkage Loss function, which correlates the compressor and its effect on reconstruction, guiding it to drop inferable tokens. SCAR achieves higher ROUGE scores on benchmark datasets than the existing state-of-the-art methods and baselines. We also conduct a user study to demonstrate the application of our model as a text highlighting system. Using our model to underscore salient information facilitates speed-reading and reduces the time required to skim a document.

Why is penguin more similar to polar bear than to sea gull? Analyzing conceptual knowledge in distributional models [Website] [PDF]
Pia Sommerauer 12:00–13:00

What do powerful models of word mean- ing created from distributional data (e.g. Word2vec (Mikolov et al., 2013) BERT (Devlin et al., 2019) and ELMO (Peters et al., 2018)) represent? What causes words to be similar in the semantic space? What type of information is lacking? This thesis proposal presents a framework for investigating the information encoded in distributional semantic models. Several analysis methods have been suggested, but they have been shown to be limited and are not well understood. This approach pairs observations made on actual corpora with insights obtained from data manipulation experiments. The expected outcome is a better understanding of (1) the semantic information we can infer purely based on linguistic co-occurrence patterns and (2) the potential of distributional semantic models to pick up linguistic evidence.

Demo Session 3B

Time: *12:45–13:30*

TextBrewer: An Open-Source Knowledge Distillation Toolkit for Natural Language Processing [Website][PDF]
Ziqing Yang, Yiming Cui, Zhipeng Chen, Wanxiang Che, Ting Liu, Shijin Wang, and Guoping Hu

In this paper, we introduce TextBrewer, an open-source knowledge distillation toolkit designed for natural language processing. It works with different neural network models and supports various kinds of supervised learning tasks, such as text classification, reading comprehension, sequence labeling. TextBrewer provides a simple and uniform workflow that enables quick setting up of distillation experiments with highly flexible configurations. It offers a set of predefined distillation methods and can be extended with custom code. As a case study, we use TextBrewer to distill BERT on several typical NLP tasks. With simple configurations, we achieve results that are comparable with or even higher than the public distilled BERT models with similar numbers of parameters.

SyntaxGym: An Online Platform for Targeted Evaluation of Language Models [Website][PDF]
Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy

Targeted syntactic evaluations have yielded insights into the generalizations learned by neural network language models. However, this line of research requires an uncommon confluence of skills: both the theoretical knowledge needed to design controlled psycholinguistic experiments, and the technical proficiency needed to train and deploy large-scale language models. We present SyntaxGym, an online platform designed to make targeted evaluations accessible to both experts in NLP and linguistics, reproducible across computing environments, and standardized following the norms of psycholinguistic experimental design. This paper releases two tools of independent value for the computational linguistics community: 1. A website, syntaxgym.org, which centralizes the process of targeted syntactic evaluation and provides easy tools for analysis and visualization; 2. Two command-line tools, 'syntaxgym' and 'lm-zoo', which allow any user to reproduce targeted syntactic evaluations and general language model inference on their own machine.

Session 3B Overview – Monday, July 6, 2020 13:00–14:00

Track A *Dialogue and Interactive Systems-6* Abstracts	Conversational Graph Grounded Policy Learning for Open-Domain Conversation Generation *Xu, Wang, Niu, Wu, Che, and Liu* [Website][PDF]	[TACL] Cross-WOZ: A Large-Scale Chinese Cross-Domain Task-Oriented Dialogue Dataset *Zhu, Huang, Zhang, Zhu, and Huang* [Website][PDF]	Dialogue State Tracking with Explicit Slot Connection Modeling *Ouyang, Chen, Dai, Zhao, Huang, and CHEN* [Website][PDF]	End-to-End Neural Pipeline for Goal-Oriented Dialogue Systems using GPT-2 *Ham, Lee, Jang, and Kim* [Website][PDF]	Gated Convolutional Bidirectional Attention-based Model for Off-topic Spoken Response Detection *Zha, Li, and Lin* [Website][PDF]
	Learning Dialog Policies from Weak Demonstrations *Gordon-Hall, Gorinski, and Cohen* [Website][PDF]	Learning to Tag OOV Tokens by Integrating Contextual Representation and Background Knowledge *He, Yan, and XU* [Website][PDF]	Multi-Agent Task-Oriented Dialog Policy Learning with Role-Aware Reward Decomposition *Takanobu, Liang, and Huang* [Website][PDF]	Semi-Supervised Dialogue Policy Learning via Stochastic Reward Estimation *Huang, Qi, Sun, and Zhang* [Website][PDF]	Slot-consistent NLG for Task-oriented Dialogue Systems with Iterative Rectification Network *Li, Yao, Qin, Che, Li, and Liu* [Website][PDF]
	Towards Conversational Recommendation over Multi-Type Dialogs *Liu, Wang, Niu, Wu, Che, and Liu* [Website][PDF]	Unknown Intent Detection Using Gaussian Mixture Model with an Application to Zero-shot Intent Classification *Yan, Fan, Li, Liu, Zhang, Wu, and Lam* [Website][PDF]			
Track B *Discourse and Pragmatics-3* Abstracts	A Complete Shift-Reduce Chinese Discourse Parser with Robust Dynamic Oracle *Hung, Huang, and Chen* [Website][PDF]	Bridging Anaphora Resolution as Question Answering *Hou* [Website][PDF]	Dialogue Coherence Assessment Without Explicit Dialogue Act Labels *Mesgar, Bücker, and Gurevych* [Website][PDF]	TransS-Driven Joint Learning Architecture for Implicit Discourse Relation Recognition *He, Wang, Guo, and Han* [Website][PDF]	
Track C *Generation-6* Abstracts	A Study of Non-autoregressive Model for Sequence Generation *Ren, Liu, Tan, Zhao, and Liu* [Website][PDF]	Expertise Style Transfer: A New Task Towards Better Communication between Experts and Laymen *Cao, Shui, Pan, Kan, Liu, and Chua* [Website][PDF]	GPT-too: A Language-Model-First Approach for AMR-to-Text Generation *Mager, Fernandez Astudillo, Naseem, Sultan, Lee, Florian, and Roukos* [Website][PDF]	Learning to Update Natural Language Comments Based on Code Changes *Panthaplackel, Nie, Gligoric, Li, and Mooney* [Website][PDF]	[TACL] Leveraging Pre-trained Checkpoints for Sequence Generation Tasks *Rothe, Narayan, and Severyn* [Website][PDF]
	Politeness Transfer: A Tag and Generate Approach *Madaan, Setlur, Parekh, Poczos, Neubig, Yang, Salakhutdinov, Black, and Prabhumoye* [Website][PDF]	Semantic Graphs for Generating Deep Questions *Pan, Xie, Feng, Chua, and Kan* [Website][PDF]	TAG : Type Auxiliary Guiding for Code Comment Generation *Cai, Liang, Xu, Hao, and Chen* [Website][PDF]	Towards Faithful Neural Table-to-Text Generation with Content-Matching Constraints *Wang, Wang, An, Yu, and Chen* [Website][PDF]	

Track D *Information Extraction-2* Abstracts	A Novel Cascade Binary Tagging Framework for Relational Triple Extraction *Wei, Su, Wang, Tian, and Chang* [Website][PDF]	In Layman's Terms: Semi-Open Relation Extraction from Scientific Texts *Kruiper, Vincent, Chen-Burger, Desmulliez, and Konstas* [Website][PDF]	NAT: Noise-Aware Training for Robust Neural Sequence Labeling *Namysl, Behnke, and Köhler* [Website][PDF]	Named Entity Recognition without Labelled Data: A Weak Supervision Approach *Lison, Barnes, Hubin, and Touileb* [Website][PDF]	Probing Linguistic Features of Sentence-Level Representations in Relation Extraction *Alt, Gabryszak, and Hennig* [Website][PDF]
	Reasoning with Latent Structure Refinement for Document-Level Relation Extraction *Nan, Guo, Sekulic, and Lu* [Website][PDF]	TACRED Revisited: A Thorough Evaluation of the TACRED Relation Extraction Task *Alt, Gabryszak, and Hennig* [Website][PDF]			
Track E *Machine Translation-4* Abstracts	BPE-Dropout: Simple and Effective Subword Regularization *Provilkov, Emelianenko, and Voita* [Website][PDF]	Bilingual Dictionary Based Neural Machine Translation without Using Parallel Sentences *Duan, Ji, Jia, Tan, Zhang, Chen, Luo, and Zhang* [Website][PDF]	Character-Level Translation with Self-attention *Gao, Nikolov, Hu, and Hahnloser* [Website][PDF]	Content Word Aware Neural Machine Translation *Chen, Wang, Utiyama, and Sumita* [Website][PDF]	Evaluating Explanation Methods for Neural Machine Translation *Li, Liu, Li, Li, Huang, and Shi* [Website][PDF]
	Improving Non-autoregressive Neural Machine Translation with Monolingual Data *Zhou and Keung* [Website][PDF]	It's Easier to Translate out of English than into it: Measuring Neural Translation Difficulty by Cross-Mutual Information *Bugliarello, Mielke, Anastasopoulos, Cotterell, and Okazaki* [Website][PDF]	Language-aware Interlingua for Multilingual Neural Machine Translation *Zhu, Yu, Cheng, and Luo* [Website][PDF]	Multiscale Collaborative Deep Models for Neural Machine Translation *Wei, Yu, Hu, Zhang, Weng, and Luo* [Website][PDF]	On the Limitations of Cross-lingual Encoders as Exposed by Reference-Free Machine Translation Evaluation *Zhao, Glavaš, Peyrard, Gao, West, and Eger* [Website][PDF]
	[CL] On the Linguistic Representational Power of Neural Machine Translation Models *Belinkov, Durrani, Dalvi, Sajjad, and Glass* [Website][PDF]	Parallel Sentence Mining by Constrained Decoding *Chen, Bogoychev, Heafield, and Kirefu* [Website][PDF]			
Track F *Phonology, Morphology and Word Segmentation-2* Abstracts	A Graph Autoencoder Model of Derivational Morphology *Hofmann, Schütze, and Pierrehumbert* [Website][PDF]				
Track G *Student Research Workshop* Abstracts	Transferring Monolingual Model to Low-Resource Language: The Case of Tigrinya *Tela, Zewoudie, and Hautamäki* [Website]	A Simple and Effective Dependency Parser for Telugu *Nallani, Shrivastava, and Sharma* [Website][PDF]	Pointwise Paraphrase Appraisal is Potentially Problematic *Chen, Ji, and Evans* [Website][PDF]	Efficient Neural Machine Translation for Low-Resource Languages via Exploiting Related Languages *Goyal, Kumar, and Sharma* [Website][PDF]	

Track H *Summarization-2* Abstracts	A Large-Scale Multi-Document Summarization Dataset from the Wikipedia Current Events Portal *Gholipour Ghalandari, Hokamp, Pham, Glover, and Ifrim* [Website][PDF]	Attend to Medical Ontologies: Content Selection for Clinical Abstractive Summarization *Sotudeh Gharebagh, Goharian, and Filice* [Website][PDF]	Attend, Translate and Summarize: An Efficient Method for Neural Cross-Lingual Summarization *Zhu, Zhou, Zhang, and Zong* [Website][PDF]	Examining the State-of-the-Art in News Timeline Summarization *Gholipour Ghalandari and Ifrim* [Website][PDF]	Improving Truthfulness of Headline Generation *Matsumaru, Takase, and Okazaki* [Website][PDF]
	On Faithfulness and Factuality in Abstractive Summarization *Maynez, Narayan, Bohnet, and McDonald* [Website][PDF]	SUPERT: Towards New Frontiers in Unsupervised Evaluation Metrics for Multi-Document Summarization *Gao, Zhao, and Eger* [Website][PDF]	Screenplay Summarization Using Latent Narrative Structure *Papalampidi, Keller, Frermann, and Lapata* [Website][PDF]	Self-Attention Guided Copy Mechanism for Abstractive Summarization *Xu, Li, Yuan, Wu, He, and Zhou* [Website][PDF]	Unsupervised Opinion Summarization with Noising and Denoising *Amplayo and Lapata* [Website][PDF]

Session 3B Details

Session 3B: Dialogue and Interactive Systems-6

Conversational Graph Grounded Policy Learning for Open-Domain Conversation Generation [Website][PDF]
Jun Xu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu 13:00–14:00

To address the challenge of policy learning in open-domain multi-turn conversation, we propose to represent prior information about dialog transitions as a graph and learn a graph grounded dialog policy, aimed at fostering a more coherent and controllable dialog. To this end, we first construct a conversational graph (CG) from dialog corpora, in which there are vertices to represent "what to say" and "how to say", and edges to represent natural transition between a message (the last utterance in a dialog context) and its response. We then present a novel CG grounded policy learning framework that conducts dialog flow planning by graph traversal, which learns to identify a what-vertex and a how-vertex from the CG at each turn to guide response generation. In this way, we effectively leverage the CG to facilitate policy learning as follows: (1) it enables more effective long-term reward design, (2) it provides high-quality candidate actions, and (3) it gives us more control over the policy. Results on two benchmark corpora demonstrate the effectiveness of this framework.

[TACL] CrossWOZ: A Large-Scale Chinese Cross-Domain Task-Oriented Dialogue Dataset [Website][PDF]
Qi Zhu, Kaili Huang, Zheng Zhang, Xiaoyan Zhu, and Minlie Huang 13:00–14:00

To advance multi-domain (cross-domain) dialogue modeling as well as alleviate the shortage of Chinese task-oriented datasets, we propose CrossWOZ, the first large-scale Chinese Cross-Domain Wizard-of-Oz task-oriented dataset. It contains 6K dialogue sessions and 102K utterances for 5 domains, including hotel, restaurant, attraction, metro, and taxi. Moreover, the corpus contains rich annotation of dialogue states and dialogue acts at both user and system sides. About 60% of the dialogues have cross-domain user goals that favor inter-domain dependency and encourage natural transition across domains in conversation. We also provide a user simulator and several benchmark models for pipelined task-oriented dialogue systems, which will facilitate researchers to compare and evaluate their models on this corpus. The large size and rich annotation of CrossWOZ make it suitable to investigate a variety of tasks in cross-domain dialogue modeling, such as dialogue state tracking, policy learning, user simulation, etc.

Dialogue State Tracking with Explicit Slot Connection Modeling [Website][PDF]
Yawen Ouyang, Moxin Chen, Xinyu Dai, Yinggong Zhao, Shujian Huang, and Jiajun CHEN 13:00–14:00

Recent proposed approaches have made promising progress in dialogue state tracking (DST). However, in multi-domain scenarios, ellipsis and reference are frequently adopted by users to express values that have been mentioned by slots from other domains. To handle these phenomena, we propose a Dialogue State Tracking with Slot Connections (DST-SC) model to explicitly consider slot correlations across different domains. Given a target slot, the slot connecting mechanism in DST-SC can infer its source slot and copy the source slot value directly, thus significantly reducing the difficulty of learning and reasoning. Experimental results verify the benefits of explicit slot connection modeling, and our model achieves state-of-the-art performance on MultiWOZ 2.0 and MultiWOZ 2.1 datasets.

End-to-End Neural Pipeline for Goal-Oriented Dialogue Systems using GPT-2 [Website][PDF]
Donghoon Ham, Jeong-Gwan Lee, Youngsoo Jang, and Kee-Eung Kim 13:00–14:00

The goal-oriented dialogue system needs to be optimized for tracking the dialogue flow and carrying out an effective conversation under various situations to meet the user goal. The traditional approach to build such a dialogue system is to take a pipelined modular architecture, where its modules are optimized individually. However, such an optimization scheme does not necessarily yield the overall performance improvement of the whole system. On the other hand, end-to-end dialogue systems with monolithic neural architecture are often trained only with input-output utterances, without taking into account the entire annotations available in the corpus. This scheme makes it difficult for goal-oriented dialogues where the system needs to integrate with external systems or to provide interpretable information about why the system generated a particular response. In this paper, we present an end-to-end neural architecture for dialogue systems that addresses both challenges above. In the human evaluation, our dialogue system achieved the success rate of 68.32%, the language understanding score of 4.149, and the response appropriateness score of 4.287, which ranked the system at the top position in the end-to-end multi-domain dialogue system task in the 8th dialogue systems technology challenge (DSTC8).

Gated Convolutional Bidirectional Attention-based Model for Off-topic Spoken Response Detection [Website][PDF]
Yefei Zha, Ruobing Li, and Hui Lin 13:00–14:00

Off-topic spoken response detection, the task aiming at predicting whether a response is off-topic for the corresponding prompt, is important for an automated speaking assessment system. In many real-world educational applications, off-topic spoken response detectors are required to achieve high recall for off-topic responses not only on seen prompts but also on prompts that are unseen during training. In this paper, we propose a novel approach for off-topic spoken response detection with high off-topic recall on both seen and unseen prompts. We introduce a new model, Gated Convolutional Bidirectional Attention-based Model (GCBiA), which applies bi-attention mechanism and convolutions to extract topic words of prompts and key-phrases of responses, and introduces gated unit and

residual connections between major layers to better represent the relevance of responses and prompts. Moreover, a new negative sampling method is proposed to augment training data. Experiment results demonstrate that our novel approach can achieve significant improvements in detecting off-topic responses with extremely high on-topic recall, for both seen and unseen prompts.

Learning Dialog Policies from Weak Demonstrations [Website][PDF]
Gabriel Gordon-Hall, Philip John Gorinski, and Shay B. Cohen 13:00–14:00

Deep reinforcement learning is a promising approach to training a dialog manager, but current methods struggle with the large state and action spaces of multi-domain dialog systems. Building upon Deep Q-learning from Demonstrations (DQfD), an algorithm that scores highly in difficult Atari games, we leverage dialog data to guide the agent to successfully respond to a user's requests. We make progressively fewer assumptions about the data needed, using labeled, reduced-labeled, and even unlabeled data to train expert demonstrators. We introduce Reinforced Fine-tune Learning, an extension to DQfD, enabling us to overcome the domain gap between the datasets and the environment. Experiments in a challenging multi-domain dialog system framework validate our approaches, and get high success rates even when trained on out-of-domain data.

Learning to Tag OOV Tokens by Integrating Contextual Representation and Background Knowledge [Website][PDF]
Keqing He, Yuanmeng Yan, and Weiran XU 13:00–14:00

Neural-based context-aware models for slot tagging have achieved state-of-the-art performance. However, the presence of OOV(out-of-vocab) words significantly degrades the performance of neural-based models, especially in a few-shot scenario. In this paper, we propose a novel knowledge-enhanced slot tagging model to integrate contextual representation of input text and the large-scale lexical background knowledge. Besides, we use multi-level graph attention to explicitly model lexical relations. The experiments show that our proposed knowledge integration mechanism achieves consistent improvements across settings with different sizes of training data on two public benchmark datasets.

Multi-Agent Task-Oriented Dialog Policy Learning with Role-Aware Reward Decomposition [Website][PDF]
Ryuichi Takanobu, Runze Liang, and Minlie Huang 13:00–14:00

Many studies have applied reinforcement learning to train a dialog policy and show great promise these years. One common approach is to employ a user simulator to obtain a large number of simulated user experiences for reinforcement learning algorithms. However, modeling a realistic user simulator is challenging. A rule-based simulator requires heavy domain expertise for complex tasks, and a data-driven simulator requires considerable data and it is even unclear how to evaluate a simulator. To avoid explicitly building a user simulator beforehand, we propose Multi-Agent Dialog Policy Learning, which regards both the system and the user as the dialog agents. Two agents interact with each other and are jointly learned simultaneously. The method uses the actor-critic framework to facilitate pre-training and improve scalability. We also propose Hybrid Value Network for the role-aware reward decomposition to integrate role-specific domain knowledge of each agent in the task-oriented dialog. Results show that our method can successfully build a system policy and a user policy simultaneously, and two agents can achieve a high task success rate through conversational interaction.

Semi-Supervised Dialogue Policy Learning via Stochastic Reward Estimation [Website][PDF]
Xinting Huang, Jianzhong Qi, Yu Sun, and Rui Zhang 13:00–14:00

Dialogue policy optimization often obtains feedback until task completion in task-oriented dialogue systems. This is insufficient for training intermediate dialogue turns since supervision signals (or rewards) are only provided at the end of dialogues. To address this issue, reward learning has been introduced to learn from state-action pairs of an optimal policy to provide turn-by-turn rewards. This approach requires complete state-action annotations of human-to-human dialogues (i.e., expert demonstrations), which is labor intensive. To overcome this limitation, we propose a novel reward learning approach for semi-supervised policy learning. The proposed approach learns a dynamics model as the reward function which models dialogue progress (i.e., state-action sequences) based on expert demonstrations, either with or without annotations. The dynamics model computes rewards by predicting whether the dialogue progress is consistent with expert demonstrations. We further propose to learn action embeddings for a better generalization of the reward function. The proposed approach outperforms competitive policy learning baselines on MultiWOZ, a benchmark multi-domain dataset.

Slot-consistent NLG for Task-oriented Dialogue Systems with Iterative Rectification Network [Website][PDF]
Yangming Li, Kaisheng Yao, Libo Qin, Wanxiang Che, Xiaolong Li, and Ting Liu 13:00–14:00

Data-driven approaches using neural networks have achieved promising performances in natural language generation (NLG). However, neural generators are prone to make mistakes, e.g., neglecting an input slot value and generating a redundant slot value. Prior works refer this to hallucination phenomenon. In this paper, we study slot consistency for building reliable NLG systems with all slot values of input dialogue act (DA) properly generated in output sentences. We propose Iterative Rectification Network (IRN) for improving general NLG systems to produce both correct and fluent responses. It applies a bootstrapping algorithm to sample training candidates and uses reinforcement learning to incorporate discrete reward related to slot inconsistency into training. Comprehensive studies have been conducted on multiple benchmark datasets, showing that the proposed methods have significantly reduced the slot error rate (ERR) for all strong baselines. Human evaluations also have confirmed its effectiveness.

Towards Conversational Recommendation over Multi-Type Dialogs [Website][PDF]
Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu 13:00–14:00

We focus on the study of conversational recommendation in the context of multi-type dialogs, where the bots can proactively and naturally lead a conversation from a non-recommendation dialog (e.g., QA) to a recommendation dialog, taking into account user's interests and feedback. To facilitate the study of this task, we create a human-to-human Chinese dialog dataset DuRecDial (about 10k dialogs, 156k utterances), where there are multiple sequential dialogs for a pair of a recommendation seeker (user) and a recommender (bot). In each dialog, the recommender proactively leads a multi-type dialog to approach recommendation targets and then makes multiple recommendations with rich interaction behavior. This dataset allows us to systematically investigate different parts of the overall problem, e.g., how to naturally lead a dialog, how to interact with users for recommendation. Finally we establish baseline results on DuRecDial for future studies.

Unknown Intent Detection Using Gaussian Mixture Model with an Application to Zero-shot Intent Classification [Website][PDF]
Guangfeng Yan, Lu Fan, Qimai Li, Han Liu, Xiaotong Zhang, Xiao-Ming Wu, and Albert Y.S. Lam 13:00–14:00

User intent classification plays a vital role in dialogue systems. Since user intent may frequently change over time in many realistic scenarios, unknown (new) intent detection has become an essential problem, where the study has just begun. This paper proposes a semantic-enhanced Gaussian mixture model (SEG) for unknown intent detection. In particular, we model utterance embeddings with a Gaussian mixture distribution and inject dynamic class semantic information into Gaussian means, which enables learning more class-concentrated embeddings that help to facilitate downstream outlier detection. Coupled with a density-based outlier detection algorithm, SEG achieves competitive results on three real task-oriented dialogue datasets in two languages for unknown intent detection. On top of that, we propose to integrate SEG as an unknown intent identifier into existing generalized zero-shot intent classification models to improve their performance. A case study on a state-of-the-art method, ReCapsNet, shows that SEG can push the classification performance to a significantly higher level.

Session 3B: Discourse and Pragmatics-3

A Complete Shift-Reduce Chinese Discourse Parser with Robust Dynamic Oracle [Website][PDF]
Shyh-Shiun Hung, Hen-Hsen Huang, and Hsin-Hsi Chen 13:00–14:00

This work proposes a standalone, complete Chinese discourse parser for practical applications. We approach Chinese discourse parsing from a variety of aspects and improve the shift-reduce parser not only by integrating the pre-trained text encoder, but also by employing novel training strategies. We revise the dynamic-oracle procedure for training the shift-reduce parser, and apply unsupervised data augmentation to enhance rhetorical relation recognition. Experimental results show that our Chinese discourse parser achieves the state-of-the-art performance.

Bridging Anaphora Resolution as Question Answering [Website][PDF]
Yufang Hou 13:00–14:00

Most previous studies on bridging anaphora resolution (Poesio et al., 2004; Hou et al., 2013b; Hou, 2018a) use the pairwise model to tackle the problem and assume that the gold mention information is given. In this paper, we cast bridging anaphora resolution as question answering based on context. This allows us to find the antecedent for a given anaphor without knowing any gold mention information (except the anaphor itself). We present a question answering framework (BARQA) for this task, which leverages the power of transfer learning. Furthermore, we propose a novel method to generate a large amount of "quasi-bridging" training data. We show that our model pre-trained on this dataset and fine-tuned on a small amount of in-domain dataset achieves new state-of-the-art results for bridging anaphora resolution on two bridging corpora (ISNotes (Markert et al., 2012) and BASHI (Rosiger, 2018)).

Dialogue Coherence Assessment Without Explicit Dialogue Act Labels [Website][PDF]
Mohsen Mesgar, Sebastian Bücker, and Iryna Gurevych 13:00–14:00

Recent dialogue coherence models use the coherence features designed for monologue texts, e.g. nominal entities, to represent utterances and then explicitly augment them with dialogue-relevant features, e.g., dialogue act labels. It indicates two drawbacks, (a) semantics of utterances are limited to entity mentions, and (b) the performance of coherence models strongly relies on the quality of the input dialogue act labels. We address these issues by introducing a novel approach to dialogue coherence assessment. We use dialogue act prediction as an auxiliary task in a multi-task learning scenario to obtain informative utterance representations for coherence assessment. Our approach alleviates the need for explicit dialogue act labels during evaluation. The results of our experiments show that our model substantially (more than 20 accuracy points) outperforms its strong competitors on the DailyDialogue corpus, and performs on par with them on the SwitchBoard corpus for ranking dialogues concerning their coherence. We release our source code.

TransS-Driven Joint Learning Architecture for Implicit Discourse Relation Recognition [Website][PDF]
Ruifang He, Jian Wang, Fengyu Guo, and Yugui Han 13:00–14:00

Implicit discourse relation recognition is a challenging task due to the lack of connectives as strong linguistic clues. Previous methods primarily encode two arguments separately or extract the specific interaction patterns for the task, which have not fully exploited the annotated relation signal. Therefore, we propose a novel TransS-driven joint learning architecture to address the issues. Specifically, based on the multi-level encoder, we 1) translate discourse relations in low-dimensional embedding space (called TransS), which could mine the latent geometric structure information of argument-relation instances; 2) further exploit the semantic features of arguments to assist discourse understanding; 3) jointly learn 1) and 2) to mutually reinforce each other to obtain the better argument representations, so as to improve the performance of the task. Extensive experimental results on the Penn Discourse TreeBank (PDTB) show that our model achieves competitive results against several state-of-the-art systems.

Session 3B: Generation-6

A Study of Non-autoregressive Model for Sequence Generation [Website][PDF]
Yi Ren, Jinglin Liu, Xu Tan, Zhou Zhao, sheng zhao sheng, and Tie-Yan Liu 13:00–14:00

Non-autoregressive (NAR) models generate all the tokens of a sequence in parallel, resulting in faster generation speed compared to their autoregressive (AR) counterparts but at the cost of lower accuracy. Different techniques including knowledge distillation and source-target alignment have been proposed to bridge the gap between AR and NAR models in various tasks such as neural machine translation (NMT), automatic speech recognition (ASR), and text to speech (TTS). With the help of those techniques, NAR models can catch up with the accuracy of AR models in some tasks but not in some others. In this work, we conduct a study to understand the difficulty of NAR sequence generation and try to answer: (1) Why NAR models can catch up with AR models in some tasks but not all? (2) Why techniques like knowledge distillation and source-target alignment can help NAR models. Since the main difference between AR and NAR models is that NAR models do not use dependency among target tokens while AR models do, intuitively the difficulty of NAR sequence generation heavily depends on the strongness of dependency among target tokens. To quantify such dependency, we propose an analysis model called CoMMA to characterize the difficulty of different NAR sequence generation tasks. We have several interesting findings: 1) Among the NMT, ASR and TTS tasks, ASR has the most target-token dependency while TTS has the least. 2) Knowledge distillation reduces the target-token dependency in target sequence and thus improves the accuracy of NAR models. 3) Source-target alignment constraint encourages dependency of a target token on source tokens and thus eases the training of NAR models.

Expertise Style Transfer: A New Task Towards Better Communication between Experts and Laymen [Website][PDF]
Yixin Cao, Ruihao Shui, Liangming Pan, Min-Yen Kan, Zhiyuan Liu, and Tat-Seng Chua 13:00–14:00

The curse of knowledge can impede communication between experts and laymen. We propose a new task of expertise style transfer and contribute a manually annotated dataset with the goal of alleviating such cognitive biases. Solving this task not only simplifies the professional language, but also improves the accuracy and expertise level of laymen descriptions using simple words. This is a challenging task, unaddressed in previous work, as it requires the models to have expert intelligence in order to modify text with a deep understanding of domain knowledge and structures. We establish the benchmark performance of five state-of-the-art models for style transfer and text simplification. The results demonstrate a significant gap between machine and human performance. We also discuss the challenges of automatic evaluation, to provide insights into future research directions. The dataset is publicly available at https://srhthu.github.io/expertise-style-transfer/.

GPT-too: A Language-Model-First Approach for AMR-to-Text Generation [Website][PDF]
Manuel Mager, Ramón Fernandez Astudillo, Tahira Naseem, Md Arafat Sultan, Young-Suk Lee, Radu Florian, and Salim Roukos 13:00–14:00

Abstract Meaning Representations (AMRs) are broad-coverage sentence-level semantic graphs. Existing approaches to generating text from AMR have focused on training sequence-to-sequence or graph-to-sequence models on AMR annotated data only. In this paper, we propose an alternative approach that combines a strong pre-trained language model with cycle consistency-based re-scoring. Despite the simplicity of the approach, our experimental results show these models outperform all previous techniques on the English LDC2017T10 dataset, including the recent use of transformer architectures. In addition to the standard evaluation metrics, we provide human evaluation experiments that further substantiate the strength of our approach.

Learning to Update Natural Language Comments Based on Code Changes [Website][PDF]
Sheena Panthaplackel, Pengyu Nie, Milos Gligoric, Junyi Jessy Li, and Raymond Mooney 13:00–14:00

We formulate the novel task of automatically updating an existing natural language comment based on changes in the body of code it accompanies. We propose an approach that learns to correlate changes across two distinct language representations, to generate a sequence of edits that are applied to the existing comment to reflect the source code modifications. We train and evaluate our model using a dataset that we collected from commit histories of open-source software projects, with each example consisting of a concurrent update to a method and its corresponding comment. We compare our approach against multiple baselines using both automatic metrics and human evaluation. Results reflect the challenge of this task and that our model outperforms baselines with respect to making edits.

[TACL] Leveraging Pre-trained Checkpoints for Sequence Generation Tasks [Website][PDF]
Sascha Rothe, Shashi Narayan, and Aliaksei Severyn 13:00–14:00

Unsupervised pre-training of large neural models has recently revolutionized Natural Language Processing. By warm-starting from the publicly released checkpoints, NLP practitioners have pushed the state-of-the-art on multiple benchmarks while saving significant amounts of compute time. So far the focus has been mainly on the Natural Language Understanding tasks. In this paper, we demonstrate the efficacy of pre-trained checkpoints for Sequence Generation. We developed a Transformer-based sequence-to-sequence model that is compatible with publicly available pre-trained BERT, GPT-2 and RoBERTa checkpoints and conducted an extensive empirical study on the utility of initializing our model, both encoder and decoder, with these checkpoints. Our models result in new state-of-the-art results on Machine Translation, Text Summarization, Sentence Splitting, and Sentence Fusion.

Politeness Transfer: A Tag and Generate Approach [Website][PDF]
Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabas Poczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W Black, and Shrimai Prabhumoye 13:00–14:00

This paper introduces a new task of politeness transfer which involves converting non-polite sentences to polite sentences while preserving the meaning. We also provide a dataset of more than 1.39 instances automatically labeled for politeness to encourage benchmark evaluations on this new task. We design a tag and generate pipeline that identifies stylistic attributes and subsequently generates a sentence in the target style while preserving most of the source content. For politeness as well as five other transfer tasks, our model outperforms the state-of-the-art methods on automatic metrics for content preservation, with a comparable or better performance on style transfer accuracy. Additionally, our model surpasses existing methods on human evaluations for grammaticality, meaning preservation and transfer accuracy across all the six style transfer tasks. The data and code is located at https://github.com/tag-and-generate.

Semantic Graphs for Generating Deep Questions [Website] [PDF]
Liangming Pan, Yuxi Xie, Yansong Feng, Tat-Seng Chua, and Min-Yen Kan 13:00–14:00

This paper proposes the problem of Deep Question Generation (DQG), which aims to generate complex questions that require reasoning over multiple pieces of information about the input passage. In order to capture the global structure of the document and facilitate reasoning, we propose a novel framework that first constructs a semantic-level graph for the input document and then encodes the semantic graph by introducing an attention-based GGNN (Att-GGNN). Afterward, we fuse the document-level and graph-level representations to perform joint training of content selection and question decoding. On the HotpotQA deep-question centric dataset, our model greatly improves performance over questions requiring reasoning over multiple facts, leading to state-of-the-art performance. The code is publicly available at https://github.com/WING-NUS/SG-Deep-Question-Generation.

TAG : Type Auxiliary Guiding for Code Comment Generation [Website] [PDF]
Ruichu Cai, Zhihao Liang, Boyan Xu, zijian li zijian, Yuexing Hao, and Yao Chen 13:00–14:00

Existing leading code comment generation approaches with the structure-to-sequence framework ignores the type information of the interpretation of the code, e.g., operator, string, etc. However, introducing the type information into the existing framework is non-trivial due to the hierarchical dependence among the type information. In order to address the issues above, we propose a Type Auxiliary Guiding encoder-decoder framework for the code comment generation task which considers the source code as an N-ary tree with type information associated with each node. Specifically, our framework is featured with a Type-associated Encoder and a Type-restricted Decoder which enables adaptive summarization of the source code. We further propose a hierarchical reinforcement learning method to resolve the training difficulties of our proposed framework. Extensive evaluations demonstrate the state-of-the-art performance of our framework with both the auto-evaluated metrics and case studies.

Towards Faithful Neural Table-to-Text Generation with Content-Matching Constraints [Website] [PDF]
Zhenyi Wang, Xiaoyang Wang, Bang An, Dong Yu, and Changyou Chen 13:00–14:00

Text generation from a knowledge base aims to translate knowledge triples to natural language descriptions. Most existing methods ignore the faithfulness between a generated text description and the original table, leading to generated information that goes beyond the content of the table. In this paper, for the first time, we propose a novel Transformer-based generation framework to achieve the goal. The core techniques in our method to enforce faithfulness include a new table-text optimal-transport matching loss and a table-text embedding similarity loss based on the Transformer model. Furthermore, to evaluate faithfulness, we propose a new automatic metric specialized to the table-to-text generation problem. We also provide detailed analysis on each component of our model in our experiments. Automatic and human evaluations show that our framework can significantly outperform state-of-the-art by a large margin.

Session 3B: Information Extraction-2

A Novel Cascade Binary Tagging Framework for Relational Triple Extraction [Website][PDF]
Zhepei Wei, Jianlin Su, Yue Wang, Yuan Tian, and Yi Chang 13:00–14:00

Extracting relational triples from unstructured text is crucial for large-scale knowledge graph construction. However, few existing works excel in solving the overlapping triple problem where multiple relational triples in the same sentence share the same entities. In this work, we introduce a fresh perspective to revisit the relational triple extraction task and propose a novel cascade binary tagging framework (CasRel) derived from a principled problem formulation. Instead of treating relations as discrete labels as in previous works, our new framework models relations as functions that map subjects to objects in a sentence, which naturally handles the overlapping problem. Experiments show that the CasRel framework already outperforms state-of-the-art methods even when its encoder module uses a randomly initialized BERT encoder, showing the power of the new tagging framework. It enjoys further performance boost when employing a pre-trained BERT encoder, outperforming the strongest baseline by 17.5 and 30.2 absolute gain in F1-score on two public datasets NYT and WebNLG, respectively. In-depth analysis on different scenarios of overlapping triples shows that the method delivers consistent performance gain across all these scenarios. The source code and data are released online.

In Layman's Terms: Semi-Open Relation Extraction from Scientific Texts [Website][PDF]
Ruben Kruiper, Julian Vincent, Jessica Chen-Burger, Marc Desmulliez, and Ioannis Konstas 13:00–14:00

Information Extraction (IE) from scientific texts can be used to guide readers to the central information in scientific documents. But narrow IE systems extract only a fraction of the information captured, and Open IE systems do not perform well on the long and complex sentences encountered in scientific texts. In this work we combine the output of both types of systems to achieve Semi-Open Relation Extraction, a new task that we explore in the Biology domain. First, we present the Focused Open Biological Information Extraction (FOBIE) dataset and use FOBIE to train a state-of-the-art narrow scientific IE system to extract trade-off relations and arguments that are central to biology texts. We then run both the narrow IE system and a state-of-the-art Open IE system on a corpus of 10K open-access scientific biological texts. We show that a significant amount (65%) of erroneous and uninformative Open IE extractions can be filtered using narrow IE extractions. Furthermore, we show that the retained extractions are significantly more often informative to a reader.

NAT: Noise-Aware Training for Robust Neural Sequence Labeling [Website][PDF]
Marcin Namysl, Sven Behnke, and Joachim Köhler 13:00–14:00

Sequence labeling systems should perform reliably not only under ideal conditions but also with corrupted inputs—as these systems often process user-generated text or follow an error-prone upstream component. To this end, we formulate the noisy sequence labeling problem, where the input may undergo an unknown noising process and propose two Noise-Aware Training (NAT) objectives that improve robustness of sequence labeling performed on perturbed input: Our data augmentation method trains a neural model using a mixture of clean and noisy samples, whereas our stability training algorithm encourages the model to create a noise-invariant latent representation. We employ a vanilla noise model at training time. For evaluation, we use both the original data and its variants perturbed with real OCR errors and misspellings. Extensive experiments on English and German named entity recognition benchmarks confirmed that NAT consistently improved robustness of popular sequence labeling models, preserving accuracy on the original input. We make our code and data publicly available for the research community.

Named Entity Recognition without Labelled Data: A Weak Supervision Approach [Website][PDF]
Pierre Lison, Jeremy Barnes, Aliaksandr Hubin, and Samia Touileb 13:00–14:00

Named Entity Recognition (NER) performance often degrades rapidly when applied to target domains that differ from the texts observed during training. When in-domain labelled data is available, transfer learning techniques can be used to adapt existing NER models to the target domain. But what should one do when there is no hand-labelled data for the target domain? This paper presents a simple but powerful approach to learn NER models in the absence of labelled data through weak supervision. The approach relies on a broad spectrum of labelling functions to automatically annotate texts from the target domain. These annotations are then merged together using a hidden Markov model which captures the varying accuracies and confusions of the labelling functions. A sequence labelling model can finally be trained on the basis of this unified annotation. We evaluate the approach on two English datasets (CoNLL 2003 and news articles from Reuters and Bloomberg) and demonstrate an improvement of about 7 percentage points in entity-level F1 scores compared to an out-of-domain neural NER model.

Probing Linguistic Features of Sentence-Level Representations in Relation Extraction [Website][PDF]
Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig 13:00–14:00

Despite the recent progress, little is known about the features captured by state-of-the-art neural relation extraction (RE) models. Common methods encode the source sentence, conditioned on the entity mentions, before classifying the relation. However, the complexity of the task makes it difficult to understand how encoder architecture and supporting linguistic knowledge affect the features learned by the encoder. We introduce 14 probing tasks targeting linguistic properties relevant to RE, and we use them to study representations learned by more than 40 different encoder architecture and linguistic feature combinations trained on two datasets, TACRED and SemEval 2010 Task 8. We find that the bias induced by the architecture and the inclusion of linguistic features are clearly expressed in the probing task performance. For example, adding contextualized word representations greatly increases performance on probing tasks with a focus on named entity and part-of-speech information, and yields better results in RE. In contrast, entity masking improves RE, but considerably lowers performance on entity type related probing tasks.

Reasoning with Latent Structure Refinement for Document-Level Relation Extraction [Website][PDF]
Guoshun Nan, Zhijiang Guo, Ivan Sekulic, and Wei Lu 13:00–14:00

Document-level relation extraction requires integrating information within and across multiple sentences of a document and capturing complex interactions between inter-sentence entities. However, effective aggregation of relevant information in the document remains a challenging research question. Existing approaches construct static document-level graphs based on syntactic trees, co-references or heuristics from the unstructured text to model the dependencies. Unlike previous methods that may not be able to capture rich non-local interactions for inference, we propose a novel model that empowers the relational reasoning across sentences by automatically inducing the latent document-level graph. We further develop a refinement strategy, which enables the model to incrementally aggregate relevant information for multi-hop reasoning. Specifically, our model achieves an F1 score of 59.05 on a large-scale document-level dataset (DocRED), significantly improving over the previous results, and also yields new state-of-the-art results on the CDR and GDA dataset. Furthermore, extensive analyses show that the model is able to discover more accurate inter-sentence relations.

TACRED Revisited: A Thorough Evaluation of the TACRED Relation Extraction Task [Website][PDF]
Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig 13:00–14:00

TACRED is one of the largest, most widely used crowdsourced datasets in Relation Extraction (RE). But, even with recent advances in unsupervised pre-training and knowledge enhanced neural RE, models still show a high error rate. In this paper, we investigate the questions: Have we reached a performance ceiling or is there still room for improvement? And how do crowd annotations, dataset, and models contribute to this error rate? To answer these questions, we first validate the most challenging 5K examples in the development and test sets using trained annotators. We find that label errors account for 8% absolute F1 test error, and that more than 50% of the examples need to be relabeled. On the relabeled test set the average F1 score of a large baseline model set improves from 62.1 to 70.1. After validation, we analyze misclassifications on the challenging instances, categorize them into linguistically motivated error groups, and verify the resulting error hypotheses on three state-of-the-art RE models. We show that two groups of ambiguous relations are responsible for most of the remaining errors and that models may adopt shallow heuristics on the dataset when entities are not masked.

Session 3B: Machine Translation-4

BPE-Dropout: Simple and Effective Subword Regularization [Website][PDF]
Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita 13:00–14:00

Subword segmentation is widely used to address the open vocabulary problem in machine translation. The dominant approach to subword segmentation is Byte Pair Encoding (BPE), which keeps the most frequent words intact while splitting the rare ones into multiple tokens. While multiple segmentations are possible even with the same vocabulary, BPE splits words into unique sequences; this may prevent a model from better learning the compositionality of words and being robust to segmentation errors. So far, the only way to overcome this BPE imperfection, its deterministic nature, was to create another subword segmentation algorithm (Kudo, 2018). In contrast, we show that BPE itself incorporates the ability to produce multiple segmentations of the same word. We introduce BPE-dropout - simple and effective subword regularization method based on and compatible with conventional BPE. It stochastically corrupts the segmentation procedure of BPE, which leads to producing multiple segmentations within the same fixed BPE framework. Using BPE-dropout during training and the standard BPE during inference improves translation quality up to 2.3 BLEU compared to BPE and up to 0.9 BLEU compared to the previous subword regularization.

Bilingual Dictionary Based Neural Machine Translation without Using Parallel Sentences [Website][PDF]
Xiangyu Duan, Baijun Ji, Hao Jia, Min Tan, Min Zhang, Boxing Chen, Weihua Luo, and Yue Zhang 13:00–14:00

In this paper, we propose a new task of machine translation (MT), which is based on no parallel sentences but can refer to a ground-truth bilingual dictionary. Motivated by the ability of a monolingual speaker learning to translate via looking up the bilingual dictionary, we propose the task to see how much potential an MT system can attain using the bilingual dictionary and large scale monolingual corpora, while is independent on parallel sentences. We propose anchored training (AT) to tackle the task. AT uses the bilingual dictionary to establish anchoring points for closing the gap between source language and target language. Experiments on various language pairs show that our approaches are significantly better than various baselines, including dictionary-based word-by-word translation, dictionary-supervised cross-lingual word embedding transformation, and unsupervised MT. On distant language pairs that are hard for unsupervised MT to perform well, AT performs remarkably better, achieving performances comparable to supervised SMT trained on more than 4M parallel sentences.

Character-Level Translation with Self-attention [Website][PDF]
Yingqiang Gao, Nikola I. Nikolov, Yuhuang Hu, and Richard H.R. Hahnloser 13:00–14:00

We explore the suitability of self-attention models for character-level neural machine translation. We test the standard transformer model, as well as a novel variant in which the encoder block combines information from nearby characters using convolutions. We perform extensive experiments on WMT and UN datasets, testing both bilingual and multilingual translation to English using up to three input languages (French, Spanish, and Chinese). Our transformer variant consistently outperforms the standard transformer at the character-level and converges faster while learning more robust character-level alignments.

Content Word Aware Neural Machine Translation [Website][PDF]
Kehai Chen, Rui Wang, Masao Utiyama, and Eiichiro Sumita 13:00–14:00

Neural machine translation (NMT) encodes the source sentence in a universal way to generate the target sentence word-by-word. However, NMT does not consider the importance of word in the sentence meaning, for example, some words (i.e., content words) express more important meaning than others (i.e., function words). To address this limitation, we first utilize word frequency information to distinguish between content and function words in a sentence, and then design a content word-aware NMT to improve translation performance. Empirical results on the WMT14 English-to-German, WMT14 English-to-French, and WMT17 Chinese-to-English translation tasks show that the proposed methods can significantly improve the performance of Transformer-based NMT.

Evaluating Explanation Methods for Neural Machine Translation [Website][PDF]
Jierui Li, Lemao Liu, Huayang Li, Guanlin Li, Guoping Huang, and Shuming Shi 13:00–14:00

Recently many efforts have been devoted to interpreting the black-box NMT models, but little progress has been made on metrics to evaluate explanation methods. Word Alignment Error Rate can be used as such a metric that matches human understanding, however, it can not measure explanation methods on those target words that are not aligned to any source word. This paper thereby makes an initial attempt to evaluate explanation methods from an alternative viewpoint. To this end, it proposes a principled metric based on fidelity in regard to the predictive behavior of the NMT model. As the exact computation for this metric is intractable, we employ an efficient approach as its approximation. On six standard translation tasks, we quantitatively evaluate several explanation methods in terms of the proposed metric and we reveal some valuable findings for these explanation methods in our experiments.

Improving Non-autoregressive Neural Machine Translation with Monolingual Data [Website][PDF]
Jiawei Zhou and Phillip Keung 13:00–14:00

Non-autoregressive (NAR) neural machine translation is usually done via knowledge distillation from an autoregressive (AR) model. Under this framework, we leverage large monolingual corpora to improve the NAR model's performance, with the goal of transferring the AR model's generalization ability while preventing overfitting. On top of a strong NAR baseline, our experimental results on the WMT14 En-De and WMT16 En-Ro news translation tasks confirm that monolingual data augmentation consistently improves the performance of the NAR model to approach the

teacher AR model's performance, yields comparable or better results than the best non-iterative NAR methods in the literature and helps reduce overfitting in the training process.

It's Easier to Translate out of English than into it: Measuring Neural Translation Difficulty by Cross-Mutual Information [Website][PDF]
Emanuele Bugliarello, Sabrina J. Mielke, Antonios Anastasopoulos, Ryan Cotterell, and Naoaki Okazaki
13:00–14:00

The performance of neural machine translation systems is commonly evaluated in terms of BLEU. However, due to its reliance on target language properties and generation, the BLEU metric does not allow an assessment of which translation directions are more difficult to model. In this paper, we propose cross-mutual information (XMI): an asymmetric information-theoretic metric of machine translation difficulty that exploits the probabilistic nature of most neural machine translation models. XMI allows us to better evaluate the difficulty of translating text into the target language while controlling for the difficulty of the target-side generation component independent of the translation task. We then present the first systematic and controlled study of cross-lingual translation difficulties using modern neural translation systems. Code for replicating our experiments is available online at https://github.com/e-bug/nmt-difficulty.

Language-aware Interlingua for Multilingual Neural Machine Translation [Website][PDF]
Changfeng Zhu, Heng Yu, Shanbo Cheng, and Weihua Luo 13:00–14:00

Multilingual neural machine translation (NMT) has led to impressive accuracy improvements in low-resource scenarios by sharing common linguistic information across languages. However, the traditional multilingual model fails to capture the diversity and specificity of different languages, resulting in inferior performance compared with individual models that are sufficiently trained. In this paper, we incorporate a language-aware interlingua into the Encoder-Decoder architecture. The interlingual network enables the model to learn a language-independent representation from the semantic spaces of different languages, while still allowing for language-specific specialization of a particular language-pair. Experiments show that our proposed method achieves remarkable improvements over state-of-the-art multilingual NMT baselines and produces comparable performance with strong individual models.

Multiscale Collaborative Deep Models for Neural Machine Translation [Website][PDF]
Xiangpeng Wei, Heng Yu, Yue Hu, Yue Zhang, Rongxiang Weng, and Weihua Luo 13:00–14:00

Recent evidence reveals that Neural Machine Translation (NMT) models with deeper neural networks can be more effective but are difficult to train. In this paper, we present a MultiScale Collaborative (MSC) framework to ease the training of NMT models that are substantially deeper than those used previously. We explicitly boost the gradient back-propagation from top to bottom levels by introducing a block-scale collaboration mechanism into deep NMT models. Then, instead of forcing the whole encoder stack directly learns a desired representation, we let each encoder block learns a fine-grained representation and enhance it by encoding spatial dependencies using a context-scale collaboration. We provide empirical evidence showing that the MSC nets are easy to optimize and can obtain improvements of translation quality from considerably increased depth. On IWSLT translation tasks with three translation directions, our extremely deep models (with 72-layer encoders) surpass strong baselines by +2.2~+3.1 BLEU points. In addition, our deep MSC achieves a BLEU score of 30.56 on WMT14 English-to-German task that significantly outperforms state-of-the-art deep NMT models. We have included the source code in supplementary materials.

On the Limitations of Cross-lingual Encoders as Exposed by Reference-Free Machine Translation Evaluation [Website][PDF]
Wei Zhao, Goran Glavaš, Maxime Peyrard, Yang Gao, Robert West, and Steffen Eger 13:00–14:00

Evaluation of cross-lingual encoders is usually performed either via zero-shot cross-lingual transfer in supervised downstream tasks or via unsupervised cross-lingual textual similarity. In this paper, we concern ourselves with reference-free machine translation (MT) evaluation where we directly compare source texts to (sometimes low-quality) system translations, which represents a natural adversarial setup for multilingual encoders. Reference-free evaluation holds the promise of web-scale comparison of MT systems. We systematically investigate a range of metrics based on state-of-the-art cross-lingual semantic representations obtained with pretrained M-BERT and LASER. We find that they perform poorly as semantic encoders for reference-free MT evaluation and identify their two key limitations, namely, (a) a semantic mismatch between representations of mutual translations and, more prominently, (b) the inability to punish "translationese", i.e., low-quality literal translations. We propose two partial remedies: (1) post-hoc re-alignment of the vector spaces and (2) coupling of semantic-similarity based metrics with target-side language modeling. In segment-level MT evaluation, our best metric surpasses reference-based BLEU by 5.7 correlation points.

[CL] On the Linguistic Representational Power of Neural Machine Translation Models [Website][PDF]
Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass 13:00–14:00

Despite the recent success of deep neural networks in natural language processing and other spheres of artificial intelligence, their interpretability remains a challenge. We analyze the representations learned by neural machine translation (NMT) models at various levels of granularity and evaluate their quality through relevant extrinsic properties. In particular, we seek answers to the following questions: (i) How accurately is word structure captured within the learned representations, which is an important aspect in translating morphologically rich languages? (ii) Do the representations capture long-range dependencies, and effectively handle syntactically divergent languages? (iii) Do the representations capture lexical semantics? We conduct a thorough investigation along several parameters: (i) Which layers in the architecture capture each of these linguistic phenomena; (ii) How does the choice of translation unit (word, character, or subword unit) impact the linguistic properties captured by the underlying representations? (iii) Do the encoder and decoder learn differently and independently? (iv) Do the representations learned by multilingual

NMT models capture the same amount of linguistic information as their bilingual counterparts? Our data-driven, quantitative evaluation illuminates important aspects in NMT models and their ability to capture various linguistic phenomena. We show that deep NMT models trained in an end-to-end fashion, without being provided any direct supervision during the training process, learn a non-trivial amount of linguistic information. Notable findings include the following observations: (i) Word morphology and part-of-speech information are captured at the lower layers of the model; (ii) In contrast, lexical semantics or non-local syntactic and semantic dependencies are better represented at the higher layers of the model; (iii) Representations learned using characters are more informed about word-morphology compared to those learned using subword units; and (iv) Representations learned by multilingual models are richer compared to bilingual models.

Parallel Sentence Mining by Constrained Decoding [Website][PDF]
Pinzhen Chen, Nikolay Bogoychev, Kenneth Heafield, and Faheem Kirefu 13:00–14:00

We present a novel method to extract parallel sentences from two monolingual corpora, using neural machine translation. Our method relies on translating sentences in one corpus, but constraining the decoding by a prefix tree built on the other corpus. We argue that a neural machine translation system by itself can be a sentence similarity scorer and it efficiently approximates pairwise comparison with a modified beam search. When benchmarked on the BUCC shared task, our method achieves results comparable to other submissions.

Session 3B: Phonology, Morphology and Word Segmentation-2

A Graph Auto-encoder Model of Derivational Morphology [Website] [PDF]
Valentin Hofmann, Hinrich Schütze, and Janet Pierrehumbert 13:00–14:00

There has been little work on modeling the morphological well-formedness (MWF) of derivatives, a problem judged to be complex and difficult in linguistics. We present a graph auto-encoder that learns embeddings capturing information about the compatibility of affixes and stems in derivation. The auto-encoder models MWF in English surprisingly well by combining syntactic and semantic information with associative information from the mental lexicon.

Session 3B: Student Research Workshop

Transferring Monolingual Model to Low-Resource Language: The Case of Tigrinya [Website]
Abrhalei Frezghi Tela, Abraham Woubie Zewoudie, and Ville Hautamäki 13:00–14:00

In recent years, transformer models have achieved great success in natural language processing tasks. Most of the current state-of-the-art NLP results are achieved by using monolingual transformer models, where the model is pre-trained using a single language unlabelled text corpus. Then, the model is fine-tuned to the specific downstream task. However, the cost of pre-training a new transformer model is high for most languages. In this work, we propose a novel transfer learning method to adopt a strong source language model, trained from a large monolingual corpus to a low-resource language. Thus, using XLNet language model, we demonstrate competitive performance with mBERT and a pre-trained target language model on the Cross-lingual Sentiment (CLS) dataset and on a new sentiment analysis dataset for low-resourced language Tigrinya. With only 10k examples of the given Tigrinya sentiment analysis dataset, English XLNet has achieved 78.88% F1-Score outperforming BERT and mBERT by 10% and 7%, respectively. More interestingly, fine-tuning (English) XLNet model on the CLS dataset has promising results compared to mBERT and even outperformed mBERT for one dataset of the Japanese language.

A Simple and Effective Dependency Parser for Telugu [Website][PDF]
Sneha Nallani, Manish Shrivastava, and Dipti Sharma 13:00–14:00

We present a simple and effective dependency parser for Telugu, a morphologically rich, free word order language. We propose to replace the rich linguistic feature templates used in the past approaches with a minimal feature function using contextual vector representations. We train a BERT model on the Telugu Wikipedia data and use vector representations from this model to train the parser. Each sentence token is associated with a vector representing the token in the context of that sentence and the feature vectors are constructed by concatenating two token representations from the stack and one from the buffer. We put the feature representations through a feedforward network and train with a greedy transition based approach. The resulting parser has a very simple architecture with minimal feature engineering and achieves state-of-the-art results for Telugu.

Pointwise Paraphrase Appraisal is Potentially Problematic [Website][PDF]
Hannah Chen, Yangfeng Ji, and David Evans 13:00–14:00

The prevailing approach for training and evaluating paraphrase identification models is constructed as a binary classification problem: the model is given a pair of sentences, and is judged by how accurately it classifies pairs as either paraphrases or non-paraphrases. This pointwise-based evaluation method does not match well the objective of most real world applications, so the goal of our work is to understand how models which perform well under pointwise evaluation may fail in practice and find better methods for evaluating paraphrase identification models. As a first step towards that goal, we show that although the standard way of fine-tuning BERT for paraphrase identification by pairing two sentences as one sequence results in a model with state-of-the-art performance, that model may perform poorly on simple tasks like identifying pairs with two identical sentences. Moreover, we show that these models may even predict a pair of randomly-selected sentences with higher paraphrase score than a pair of identical ones.

Efficient Neural Machine Translation for Low-Resource Languages via Exploiting Related Languages
[Website][PDF]
Vikrant Goyal, Sourav Kumar, and Dipti Misra Sharma 13:00–14:00

A large percentage of the world's population speaks a language of the Indian subcontinent, comprising languages from both Indo-Aryan (e.g. Hindi, Punjabi, Gujarati, etc.) and Dravidian (e.g. Tamil, Telugu, Malayalam, etc.) families. A universal characteristic of Indian languages is their complex morphology, which, when combined with the general lack of sufficient quantities of high-quality parallel data, can make developing machine translation (MT) systems for these languages difficult. Neural Machine Translation (NMT) is a rapidly advancing MT paradigm and has shown promising results for many language pairs, especially in large training data scenarios. Since the condition of large parallel corpora is not met for Indian-English language pairs, we present our efforts towards building efficient NMT systems between Indian languages (specifically Indo-Aryan languages) and English via efficiently exploiting parallel data from the related languages. We propose a technique called Unified Transliteration and Subword Segmentation to leverage language similarity while exploiting parallel data from related language pairs. We also propose a Multilingual Transfer Learning technique to leverage parallel data from multiple related languages to assist translation for low resource language pair of interest. Our experiments demonstrate an overall average improvement of 5 BLEU points over the standard Transformer-based NMT baselines.

Session 3B: Summarization-2

A Large-Scale Multi-Document Summarization Dataset from the Wikipedia Current Events Portal [Website][PDF]
Demian Gholipour Ghalandari, Chris Hokamp, Nghia The Pham, John Glover, and Georgiana Ifrim 13:00–14:00

Multi-document summarization (MDS) aims to compress the content in large document collections into short summaries and has important applications in story clustering for newsfeeds, presentation of search results, and timeline generation. However, there is a lack of datasets that realistically address such use cases at a scale large enough for training supervised models for this task. This work presents a new dataset for MDS that is large both in the total number of document clusters and in the size of individual clusters. We build this dataset by leveraging the Wikipedia Current Events Portal (WCEP), which provides concise and neutral human-written summaries of news events, with links to external source articles. We also automatically extend these source articles by looking for related articles in the Common Crawl archive. We provide a quantitative analysis of the dataset and empirical results for several state-of-the-art MDS techniques.

Attend to Medical Ontologies: Content Selection for Clinical Abstractive Summarization [Website][PDF]
Sajad Sotudeh Gharebagh, Nazli Goharian, and Ross Filice 13:00–14:00

Sequence-to-sequence (seq2seq) network is a well-established model for text summarization task. It can learn to produce readable content; however, it falls short in effectively identifying key regions of the source. In this paper, we approach the content selection problem for clinical abstractive summarization by augmenting salient ontological terms into the summarizer. Our experiments on two publicly available clinical data sets (107,372 reports of MIMIC-CXR, and 3,366 reports of OpenI) show that our model statistically significantly boosts state-of-the-art results in terms of ROUGE metrics (with improvements: 2.9% RG-1, 2.5% RG-2, 1.9% RG-L), in the healthcare domain where any range of improvement impacts patients' welfare.

Attend, Translate and Summarize: An Efficient Method for Neural Cross-Lingual Summarization [Website][PDF]
Junnan Zhu, Yu Zhou, Jiajun Zhang, and Chengqing Zong 13:00–14:00

Cross-lingual summarization aims at summarizing a document in one language (e.g., Chinese) into another language (e.g., English). In this paper, we propose a novel method inspired by the translation pattern in the process of obtaining a cross-lingual summary. We first attend to some words in the source text, then translate them into the target language, and summarize to get the final summary. Specifically, we first employ the encoder-decoder attention distribution to attend to the source words. Second, we present three strategies to acquire the translation probability, which helps obtain the translation candidates for each source word. Finally, each summary word is generated either from the neural distribution or from the translation candidates of source words. Experimental results on Chinese-to-English and English-to-Chinese summarization tasks have shown that our proposed method can significantly outperform the baselines, achieving comparable performance with the state-of-the-art.

Examining the State-of-the-Art in News Timeline Summarization [Website][PDF]
Demian Gholipour Ghalandari and Georgiana Ifrim 13:00–14:00

Previous work on automatic news timeline summarization (TLS) leaves an unclear picture about how this task can generally be approached and how well it is currently solved. This is mostly due to the focus on individual subtasks, such as date selection and date summarization, and to the previous lack of appropriate evaluation metrics for the full TLS task. In this paper, we compare different TLS strategies using appropriate evaluation frameworks, and propose a simple and effective combination of methods that improves over the stateof-the-art on all tested benchmarks. For a more robust evaluation, we also present a new TLS dataset, which is larger and spans longer time periods than previous datasets.

Improving Truthfulness of Headline Generation [Website][PDF]
Kazuki Matsumaru, Sho Takase, and Naoaki Okazaki 13:00–14:00

Most studies on abstractive summarization report ROUGE scores between system and reference summaries. However, we have a concern about the truthfulness of generated summaries: whether all facts of a generated summary are mentioned in the source text. This paper explores improving the truthfulness in headline generation on two popular datasets. Analyzing headlines generated by the state-of-the-art encoder-decoder model, we show that the model sometimes generates untruthful headlines. We conjecture that one of the reasons lies in untruthful supervision data used for training the model. In order to quantify the truthfulness of article-headline pairs, we consider the textual entailment of whether an article entails its headline. After confirming quite a few untruthful instances in the datasets, this study hypothesizes that removing untruthful instances from the supervision data may remedy the problem of the untruthful behaviors of the model. Building a binary classifier that predicts an entailment relation between an article and its headline, we filter out untruthful instances from the supervision data. Experimental results demonstrate that the headline generation model trained on filtered supervision data shows no clear difference in ROUGE scores but remarkable improvements in automatic and manual evaluations of the generated headlines.

On Faithfulness and Factuality in Abstractive Summarization [Website][PDF]
Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald 13:00–14:00

It is well known that the standard likelihood training and approximate decoding objectives in neural text generation models lead to less human-like responses for open-ended tasks such as language modeling and story generation. In this paper we have analyzed limitations of these models for abstractive document summarization and found that these models are highly prone to hallucinate content that is unfaithful to the input document. We conducted a large scale human evaluation of several neural abstractive summarization systems to better understand the types of hallucinations they produce. Our human annotators found substantial amounts of hallucinated content in all model generated summaries. However, our analysis does show that pretrained models are better summarizers not only in terms of raw metrics, i.e., ROUGE, but also in generating faithful and factual summaries as evaluated by humans. Furthermore, we show that textual entailment measures better correlate with faithfulness than standard metrics, potentially leading the way to automatic evaluation metrics as well as training and decoding criteria.

SUPERT: Towards New Frontiers in Unsupervised Evaluation Metrics for Multi-Document Summarization [Website][PDF]
Yang Gao, Wei Zhao, and Steffen Eger 13:00–14:00

We study unsupervised multi-document summarization evaluation metrics, which require neither human-written reference summaries nor human annotations (e.g. preferences, ratings, etc.). We propose SUPERT, which rates the quality of a summary by measuring its semantic similarity with a pseudo reference summary, i.e. selected salient sentences from the source documents, using contextualized embeddings and soft token alignment techniques. Compared to the state-of-the-art unsupervised evaluation metrics, SUPERT correlates better with human ratings by 18-39%. Furthermore, we use SUPERT as rewards to guide a neural-based reinforcement learning summarizer, yielding favorable performance compared to the state-of-the-art unsupervised summarizers. All source code is available at https://github.com/yg211/acl20-ref-free-eval.

Screenplay Summarization Using Latent Narrative Structure [Website][PDF]
Pinelopi Papalampidi, Frank Keller, Lea Frermann, and Mirella Lapata 13:00–14:00

Most general-purpose extractive summarization models are trained on news articles, which are short and present all important information upfront. As a result, such models are biased on position and often perform a smart selection of sentences from the beginning of the document. When summarizing long narratives, which have complex structure and present information piecemeal, simple position heuristics are not sufficient. In this paper, we propose to explicitly incorporate the underlying structure of narratives into general unsupervised and supervised extractive summarization models. We formalize narrative structure in terms of key narrative events (turning points) and treat it as latent in order to summarize screenplays (i.e., extract an optimal sequence of scenes). Experimental results on the CSI corpus of TV screenplays, which we augment with scene-level summarization labels, show that latent turning points correlate with important aspects of a CSI episode and improve summarization performance over general extractive algorithms leading to more complete and diverse summaries.

Self-Attention Guided Copy Mechanism for Abstractive Summarization [Website][PDF]
Song Xu, Haoran Li, Peng Yuan, Youzheng Wu, Xiaodong He, and Bowen Zhou 13:00–14:00

Copy module has been widely equipped in the recent abstractive summarization models, which facilitates the decoder to extract words from the source into the summary. Generally, the encoder-decoder attention is served as the copy distribution, while how to guarantee that important words in the source are copied remains a challenge. In this work, we propose a Transformer-based model to enhance the copy mechanism. Specifically, we identify the importance of each source word based on the degree centrality with a directed graph built by the self-attention layer in the Transformer. We use the centrality of each source word to guide the copy process explicitly. Experimental results show that the self-attention graph provides useful guidance for the copy distribution. Our proposed models significantly outperform the baseline methods on the CNN/Daily Mail dataset and the Gigaword dataset.

Unsupervised Opinion Summarization with Noising and Denoising [Website][PDF]
Reinald Kim Amplayo and Mirella Lapata 13:00–14:00

The supervised training of high-capacity models on large datasets containing hundreds of thousands of document-summary pairs is critical to the recent success of deep learning techniques for abstractive summarization. Unfortunately, in most domains (other than news) such training data is not available and cannot be easily sourced. In this paper we enable the use of supervised learning for the setting where there are only documents available (e.g., product or business reviews) without ground truth summaries. We create a synthetic dataset from a corpus of user reviews by sampling a review, pretending it is a summary, and generating noisy versions thereof which we treat as pseudo-review input. We introduce several linguistically motivated noise generation functions and a summarization model which learns to denoise the input and generate the original review. At test time, the model accepts genuine reviews and generates a summary containing salient opinions, treating those that do not reach consensus as noise. Extensive automatic and human evaluation shows that our model brings substantial improvements over both abstractive and extractive baselines.

Demo Session 3C

Time: *13:30–14:15*

Tabouid: a Wikipedia-based word guessing game [Website][PDF]
Timothée Bernard

We present Tabouid, a word-guessing game automatically generated from Wikipedia. Tabouid contains 10,000 (virtual) cards in English, and as many in French, covering not only words and linguistic expressions but also a variety of topics including artists, historical events or scientific concepts. Each card corresponds to a Wikipedia article, and conversely, any article could be turned into a card. A range of relatively simple NLP and machine-learning techniques are effectively integrated into a two-stage process. First, a large subset of Wikipedia articles are scored - this score estimates the difficulty, or alternatively, the playability of the page. Then, the best articles are turned into cards by selecting, for each of them, a list of banned words based on its content. We believe that the game we present is more than mere entertainment and that, furthermore, this paper has pedagogical potential.

Syntactic Search by Example [Website][PDF]
Micah Shlain, Hillel Taub-Tabib, Shoval Sadde, and Yoav Goldberg

We present a system that allows a user to search a large linguistically annotated corpus using syntactic patterns over dependency graphs. In contrast to previous attempts to this effect, we introduce a light-weight query language that does not require the user to know the details of the underlying syntactic representations, and instead to query the corpus by providing an example sentence coupled with simple markup. Search is performed at an interactive speed due to efficient linguistic graph-indexing and retrieval engine. This allows for rapid exploration, development and refinement of syntax-based queries. We demonstrate the system using queries over two corpora: the English wikipedia, and a collection of English pubmed abstracts. A demo of the wikipedia system is available at https://allenai.github.io/spike/ .

Demo Session 4A

Time: *17:00–17:45*

Personalized PageRank with Syntagmatic Information for Multilingual Word Sense Disambiguation
[Website][PDF]
Federico Scozzafava, Marco Maru, Fabrizio Brignone, Giovanni Torrisi, and Roberto Navigli

Exploiting syntagmatic information is an encouraging research focus to be pursued in an effort to close the gap between knowledge-based and supervised Word Sense Disambiguation (WSD) performance. We follow this direction in our next-generation knowledge-based WSD system, SyntagRank, which we make available via a Web interface and a RESTful API. SyntagRank leverages the disambiguated pairs of co-occurring words included in SyntagNet, a lexical-semantic combination resource, to perform state-of-the-art knowledge-based WSD in a multilingual setting. Our service provides both a user-friendly interface, available at http://syntagnet.org/, and a RESTful endpoint to query the system programmatically (accessible at http://api.syntagnet.org/).

GAIA: A Fine-grained Multimedia Knowledge Extraction System [Website][PDF]
Manling Li, Alireza Zareian, Ying Lin, Xiaoman Pan, Spencer Whitehead, Brian Chen, Bo Wu, Heng Ji, Shih-Fu Chang, Clare Voss, Daniel Napierski, and Marjorie Freedman

We present the first comprehensive, open source multimedia knowledge extraction system that takes a massive stream of unstructured, heterogeneous multimedia data from various sources and languages as input, and creates a coherent, structured knowledge base, indexing entities, relations, and events, following a rich, fine-grained ontology. Our system, GAIA, enables seamless search of complex graph queries, and retrieves multimedia evidence including text, images and videos. GAIA achieves top performance at the recent NIST TAC SM-KBP2019 evaluation. The system is publicly available at GitHub and DockerHub, with a narrated video that documents the system.

Multilingual Universal Sentence Encoder for Semantic Retrieval [Website][PDF]
Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil

We present easy-to-use retrieval focused multilingual sentence embedding models, made available on TensorFlow Hub. The models embed text from 16 languages into a shared semantic space using a multi-task trained dual-encoder that learns tied cross-lingual representations via translation bridge tasks (Chidambaram et al., 2018). The models achieve a new state-of-the-art in performance on monolingual and cross-lingual semantic retrieval (SR). Competitive performance is obtained on the related tasks of translation pair bitext retrieval (BR) and retrieval question answering (ReQA). On transfer learning tasks, our multilingual embeddings approach, and in some cases exceed, the performance of English only sentence embeddings.

Keynote Address: Kathleen R. McKeown

Rewriting the Past: Assessing the Field through the Lens of Language Generation

Abstract: In recent years, we have seen tremendous advances in the field of natural language processing through the use of neural networks. In fact, they have done so well, that they have almost succeeded in rewriting the field as we knew it. In this talk, I examine the state of the field and its link to the past, with a focus on language generation of many forms. I ask where neural networks have been particularly successful, where approaches from the past might still be valuable, and where we need to turn in the future if we are to go beyond our current success. To answer these questions, this talk will feature clips from a series of interviews I carried out with experts in the field.

Biography: Kathleen R. McKeown is the Henry and Gertrude Rothschild Professor of Computer Science at Columbia University and the Founding Director of the Data Science Institute, serving as Director from 2012 to 2017. She is also an Amazon Scholar. In earlier years, she served as Department Chair 1(998-2003) and as Vice Dean for Research for the School of Engineering and Applied Science (2010-2012). A leading scholar and researcher in the field of natural language processing, McKeown focuses her research on the use of data for societal problems; her interests include text summarization, question answering, natural language generation, social media analysis and multilingual applications. She has received numerous honors and awards, including American Academy of Arts and Science elected member, American Association of Artificial Intelligence Fellow, a Founding Fellow of the Association for Computational Linguistics and an Association for Computing Machinery Fellow. Early on she received the National Science Foundation Presidential Young Investigator Award, and a National Science Foundation Faculty Award for Women. In 2010, she won both the Columbia Great Teacher Award—an honor bestowed by the students—and the Anita Borg Woman of Vision Award for Innovation.

Website: `http://www1.cs.columbia.edu/ kathy/`

Session 4A Overview – Monday, July 6, 2020 17:00–18:00

Track A *Cognitive Modeling and Psycholinguistics-4* Abstracts	A Tale of Two Perplexities: Sensitivity of Neural Language Models to Lexical Retrieval Deficits in Dementia of the Alzheimer's Type *Cohen and Pakhomov* [Website][PDF]	Inflecting When There's No Majority: Limitations of Encoder-Decoder Neural Networks as Cognitive Models for German Plurals *McCurdy, Goldwater, and Lopez* [Website][PDF]	Probing Linguistic Systematicity *Goodwin, Sinha, and O'Donnell* [Website][PDF]	Recollection versus Imagination: Exploring Human Memory and Cognition via Neural Language Models *Sap, horvitz, Choi, Smith, and Pennebaker* [Website][PDF]	Recurrent Neural Network Language Models Always Learn English-Like Relative Clause Attachment *Davis and Schijndel* [Website][PDF]
	Speakers enhance contextually confusable words *Meinhardt, Bakovic, and Bergen* [Website][PDF]	What determines the order of adjectives in English? Comparing efficiency-based theories using dependency treebanks *Futrell, Dyer, and Scontras* [Website][PDF]	You Don't Have Time to Read This: An Exploration of Document Reading Time Prediction *Weller, Hildebrandt, Reznik, Challis, Tass, Snell, and Seppi* [Website][PDF]		
Track B *Dialogue and Interactive Systems-7* Abstracts	"None of the Above": Measure Uncertainty in Dialog Response Retrieval *Feng, Mehri, Eskenazi, and Zhao* [Website][PDF]	A Generative Model for Joint Natural Language Understanding and Generation *Tseng, Cheng, Fang, and Vandyke* [Website][PDF]	Can You Put it All Together: Evaluating Conversational Agents' Ability to Blend Skills *Smith, Williamson, Shuster, Weston, and Boureau* [Website][PDF]	Efficient Dialogue State Tracking by Selectively Overwriting Memory *Kim, Yang, Kim, and Lee* [Website][PDF]	Grounded Conversation Generation as Guided Traverses in Commonsense Knowledge Graphs *Zhang, Liu, Xiong, and Liu* [Website][PDF]
	Negative Training for Neural Dialogue Response Generation *He and Glass* [Website][PDF]	Recursive Template-based Frame Generation for Task Oriented Dialog *Gangadharaiah and Narayanaswamy* [Website][PDF]	Speak to your Parser: Interactive Text-to-SQL with Natural Language Feedback *Elgohary, Hosseini, and Hassan Awadallah* [Website][PDF]	[CL] The Design and Implementation of XiaoIce, an Empathetic Social Chatbot *Zhou, Gao, Li, and Shum* [Website][PDF]	USR: An Unsupervised and Reference Free Evaluation Metric for Dialog Generation *Mehri and Eskenazi* [Website][PDF]
	Zero-Shot Transfer Learning with Synthesized Data for Multi-Domain Dialogue State Tracking *Campagna, Foryciarz, Moradshahi, and Lam* [Website][PDF]				
Track C *Machine Learning for NLP-1* Abstracts	Calibrating Structured Output Predictors for Natural Language Processing *Jagannatha and* [Website][PDF]	Active Imitation Learning with Noisy Guidance *Brantley, Daumé III, and Sharaf* [Website][PDF]	ExpBERT: Representation Engineering with Natural Language Explanations *Murty, Koh, and Liang* [Website][PDF]	GAN-BERT: Generative Adversarial Learning for Robust Text Classification with a Bunch of Labeled Examples *Croce, Castellucci, and Basili* [Website][PDF]	Generalizing Natural Language Analysis through Span-relation Representations *Jiang, Xu, Araki, and Neubig* [Website][PDF]

	Learning to Contextually Aggregate Multi-Source Supervision for Sequence Labeling *Lan, Huang, Lin, Jiang, Liu, and Ren* [Website][PDF]	MixText: Linguistically-Informed Interpolation of Hidden Space for Semi-Supervised Text Classification *Chen, Yang, and Yang* [Website][PDF]	MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices *Sun, Yu, Song, Liu, Yang, and Zhou* [Website][PDF]	On Importance Sampling-Based Evaluation of Latent Language Models *Logan IV, Gardner, and Singh* [Website][PDF]	SMART: Robust and Efficient Fine-Tuning for Pre-trained Natural Language Models through Principled Regularized Optimization *Jiang, He, Chen, Liu, Gao, and Zhao* [Website][PDF]
	Stolen Probability: A Structural Weakness of Neural Language Models *Demeter, Kimmel, and Downey* [Website][PDF]	Taxonomy Construction of Unseen Domains via Graph-based Cross-Domain Knowledge Transfer *Shang, Dash, Chowdhury, Mihindukulasooriya, and Gliozzo* [Website][PDF]	To Pretrain or Not to Pretrain: Examining the Benefits of Pretrainng on Resource Rich Tasks *Wang, Khabsa, and Ma* [Website][PDF]	Why Overfitting Isn't Always Bad: Retrofitting Cross-Lingual Word Embeddings to Dictionaries *Zhang, Fujinuma, Paul, and Boyd-Graber* [Website][PDF]	XtremeDistil: Multi-stage Distillation for Massive Multilingual Models *Mukherjee and Hassan Awadallah* [Website][PDF]
Track D *NLP Applications-3* Abstracts	A Girl Has A Name: Detecting Authorship Obfuscation *Mahmood, Shafiq, and Srinivasan* [Website][PDF]	DeeBERT: Dynamic Early Exiting for Accelerating BERT Inference *Xin, Tang, Lee, Yu, and Lin* [Website][PDF]	Efficient Strategies for Hierarchical Text Classification: External Knowledge and Auxiliary Tasks *Rivas Rojas, Bustamante, Oncevay, and Sobrevilla Cabezudo* [Website][PDF]	Investigating the effect of auxiliary objectives for the automated grading of learner English speech transcriptions *Craighead, Caines, Buttery, and Yannakoudakis* [Website][PDF]	MMPE: A Multi-Modal Interface for Post-Editing Machine Translation *Herbig, Düwel, Pal, Meladaki, Monshizadeh, Krüger, and Genabith* [Website][PDF]
	SPECTER: Document-level Representation Learning using Citation-informed Transformers *Cohan, Feldman, Beltagy, Downey, and Weld* [Website][PDF]	Semantic Scaffolds for Pseudocode-to-Code Generation *Zhong, Stern, and Klein* [Website][PDF]			
Track E *Question Answering-3* Abstracts	Contextualized Sparse Representations for Real-Time Open-Domain Question Answering *Lee, Seo, Hajishirzi, and Kang* [Website][PDF]	Dynamic Sampling Strategies for Multi-Task Reading Comprehension *Gottumukkala, Dua, Singh, and Gardner* [Website][PDF]			
Track F *Textual Inference and Other Areas of Semantics-1* Abstracts	Can We Predict New Facts with Open Knowledge Graph Embeddings? A Benchmark for Open Link Prediction *Broscheit, Gashteovski, Wang, and Gemulla* [Website][PDF]	[TACL] Decomposing Generalization: Models of Generic, Habitual and Episodic Statements *Govindarajan, Durme, and White* [Website][PDF]	INFOTABS: Inference on Tables as Semi-structured Data *Gupta, Mehta, Nokhiz, and Srikumar* [Website][PDF]	[TACL] Inherent Disagreements in Human Textual Inferences *Pavlick and Kwiatkowski* [Website][PDF]	Interactive Machine Comprehension with Information Seeking Agents *Yuan, Fu, Côté, Tay, Pal, and Trischler* [Website][PDF]

	Syntactic Data Augmentation Increases Robustness to Inference Heuristics *Min, McCoy, Das, Pitler, and Linzen* [Website][PDF]				
Track G *Speech and Multimodality-1* Abstracts	[TACL] Acoustic-Prosodic and Lexical Cues to Deception and Trust: Deciphering How People Detect Lies *Chen, Levitan, Levine, Mandic, and Hirschberg* [Website][PDF]	Improved Speech Representations with Multi-Target Autoregressive Predictive Coding *Chung and Glass* [Website][PDF]	Integrating Multimodal Information in Large Pretrained Transformers *Rahman, Hasan, Lee, Bagher Zadeh, Mao, Morency, and Hoque* [Website][PDF]	MultiQT: Multimodal learning for real-time question tracking in speech *D. Havtorn, Latko, Edin, Maaløe, Borgholt, Belgrano, Jacobsen, Sdun, and Agić* [Website][PDF]	Multimodal and Multiresolution Speech Recognition with Transformers *Paraskevopoulos, Parthasarathy, Khare, and Sundaram* [Website][PDF]
	Phone Features Improve Speech Translation *Salesky and Black* [Website][PDF]				
Track H *Student Research Workshop* Abstracts	Media Bias, the Social Sciences, and NLP: Automating Frame Analyses to Identify Bias by Word Choice and Labeling *Hamborg* [Website][PDF]	Exploring Interpretability in Event Extraction: Multitask Learning of a Neural Event Classifier and an Explanation Decoder *Tang, Hahn-Powell, and Surdeanu* [Website][PDF]	Research Replication Prediction Using Weakly Supervised Learning *Luo, Li, Wang, and Liu* [Website]	Crossing the Line: Where do Demographic Variables Fit into Humor Detection? *Meaney* [Website][PDF]	

Session 4A Details

Session 4A: Cognitive Modeling and Psycholinguistics-4

A Tale of Two Perplexities: Sensitivity of Neural Language Models to Lexical Retrieval Deficits in Dementia of the Alzheimer's Type [Website][PDF]
Trevor Cohen and Serguei Pakhomov 17:00–18:00

In recent years there has been a burgeoning interest in the use of computational methods to distinguish between elicited speech samples produced by patients with dementia, and those from healthy controls. The difference between perplexity estimates from two neural language models (LMs) - one trained on transcripts of speech produced by healthy participants and one trained on those with dementia - as a single feature for diagnostic classification of unseen transcripts has been shown to produce state-of-the-art performance. However, little is known about why this approach is effective, and on account of the lack of case/control matching in the most widely-used evaluation set of transcripts (DementiaBank), it is unclear if these approaches are truly diagnostic, or are sensitive to other variables. In this paper, we interrogate neural LMs trained on participants with and without dementia by using synthetic narratives previously developed to simulate progressive semantic dementia by manipulating lexical frequency. We find that perplexity of neural LMs is strongly and differentially associated with lexical frequency, and that using a mixture model resulting from interpolating control and dementia LMs improves upon the current state-of-the-art for models trained on transcript text exclusively.

Inflecting When There's No Majority: Limitations of Encoder-Decoder Neural Networks as Cognitive Models for German Plurals [Website][PDF]
Kate McCurdy, Sharon Goldwater, and Adam Lopez 17:00–18:00

Can artificial neural networks learn to represent inflectional morphology and generalize to new words as human speakers do? Kirov and Cotterell (2018) argue that the answer is yes: modern Encoder-Decoder (ED) architectures learn human-like behavior when inflecting English verbs, such as extending the regular past tense form /-(e)d/ to novel words. However, their work does not address the criticism raised by Marcus et al. (1995): that neural models may learn to extend not the regular, but the most frequent class — and thus fail on tasks like German number inflection, where infrequent suffixes like /-s/ can still be productively generalized. To investigate this question, we first collect a new dataset from German speakers (production and ratings of plural forms for novel nouns) that is designed to avoid sources of information unavailable to the ED model. The speaker data show high variability, and two suffixes evince 'regular' behavior, appearing more often with phonologically atypical inputs. Encoder-decoder models do generalize the most frequently produced plural class, but do not show human-like variability or 'regular' extension of these other plural markers. We conclude that modern neural models may still struggle with minority-class generalization.

Probing Linguistic Systematicity [Website][PDF]
Emily Goodwin, Koustuv Sinha, and Timothy J. O'Donnell 17:00–18:00

Recently, there has been much interest in the question of whether deep natural language understanding (NLU) models exhibit systematicity, generalizing such that units like words make consistent contributions to the meaning of the sentences in which they appear. There is accumulating evidence that neural models do not learn systematically. We examine the notion of systematicity from a linguistic perspective, defining a set of probing tasks and a set of metrics to measure systematic behaviour. We also identify ways in which network architectures can generalize non-systematically, and discuss why such forms of generalization may be unsatisfying. As a case study, we perform a series of experiments in the setting of natural language inference (NLI). We provide evidence that current state-of-the-art NLU systems do not generalize systematically, despite overall high performance.

Recollection versus Imagination: Exploring Human Memory and Cognition via Neural Language Models [Website][PDF]
Maarten Sap, Eric horvitz, Yejin Choi, Noah A. Smith, and James Pennebaker 17:00–18:00

We investigate the use of NLP as a measure of the cognitive processes involved in storytelling, contrasting imagination and recollection of events. To facilitate this, we collect and release Hippocorpus, a dataset of 7,000 stories about imagined and recalled events. We introduce a measure of narrative flow and use this to examine the narratives for imagined and recalled events. Additionally, we measure the differential recruitment of knowledge attributed to semantic memory versus episodic memory (Tulving, 1972) for imagined and recalled storytelling by comparing the frequency of descriptions of general commonsense events with more specific realis events. Our analyses show that imagined stories have a substantially more linear narrative flow, compared to recalled stories in which adjacent sentences are more disconnected. In addition, while recalled stories rely more on autobiographical events based on episodic memory, imagined stories express more commonsense knowledge based on semantic memory. Finally, our measures reveal the effect of narrativization of memories in stories (e.g., stories about frequently recalled memories flow more linearly; Bartlett, 1932). Our findings highlight the potential of using NLP tools to study the traces of human cognition in language.

Recurrent Neural Network Language Models Always Learn English-Like Relative Clause Attachment [Website][PDF]
Forrest Davis and Marten van Schijndel 17:00–18:00

A standard approach to evaluating language models analyzes how models assign probabilities to valid versus invalid syntactic constructions (i.e. is a grammatical sentence more probable than an ungrammatical sentence). Our work uses ambiguous relative clause attachment to extend such evaluations to cases of multiple simultaneous valid interpretations, where stark grammaticality differences are absent. We compare model performance in English and Spanish to show that non-linguistic biases in RNN LMs advantageously overlap with syntactic structure in English but not Spanish. Thus, English models may appear to acquire human-like syntactic preferences, while models trained on Spanish fail to acquire comparable human-like preferences. We conclude by relating these results to broader concerns about the relationship between comprehension (i.e. typical language model use cases) and production (which generates the training data for language models), suggesting that necessary linguistic biases are not present in the training signal at all.

Speakers enhance contextually confusable words [Website][PDF]
Eric Meinhardt, Eric Bakovic, and Leon Bergen 17:00–18:00

Recent work has found evidence that natural languages are shaped by pressures for efficient communication — e.g. the more contextually predictable a word is, the fewer speech sounds or syllables it has (Piantadosi et al. 2011). Research on the degree to which speech and language are shaped by pressures for effective communication — robustness in the face of noise and uncertainty — has been more equivocal. We develop a measure of contextual confusability during word recognition based on psychoacoustic data. Applying this measure to naturalistic speech corpora, we find evidence suggesting that speakers alter their productions to make contextually more confusable words easier to understand.

What determines the order of adjectives in English? Comparing efficiency-based theories using dependency treebanks [Website][PDF]
Richard Futrell, William Dyer, and Greg Scontras 17:00–18:00

We take up the scientific question of what determines the preferred order of adjectives in English, in phrases such as big blue box where multiple adjectives modify a following noun. We implement and test four quantitative theories, all of which are theoretically motivated in terms of efficiency in human language production and comprehension. The four theories we test are subjectivity (Scontras et al., 2017), information locality (Futrell, 2019), integration cost (Dyer, 2017), and information gain, which we introduce. We evaluate theories based on their ability to predict orders of unseen adjectives in hand-parsed and automatically-parsed dependency treebanks. We find that subjectivity, information locality, and information gain are all strong predictors, with some evidence for a two-factor account, where subjectivity and information gain reflect a factor involving semantics, and information locality reflects collocational preferences.

You Don't Have Time to Read This: An Exploration of Document Reading Time Prediction [Website][PDF]
Orion Weller, Jordan Hildebrandt, Ilya Reznik, Christopher Challis, E. Shannon Tass, Quinn Snell, and Kevin Seppi 17:00–18:00

Predicting reading time has been a subject of much previous work, focusing on how different words affect human processing, measured by reading time. However, previous work has dealt with a limited number of participants as well as word level only predictions (i.e. predicting the time to read a single word). We seek to extend these works by examining whether or not document level predictions are effective, given additional information such as subject matter, font characteristics, and readability metrics. We perform a novel experiment to examine how different features of text contribute to the time it takes to read, distributing and collecting data from over a thousand participants. We then employ a large number of machine learning methods to predict a user's reading time. We find that despite extensive research showing that word level reading time can be most effectively predicted by neural networks, larger scale text can be easily and most accurately predicted by one factor, the number of words.

Session 4A: Dialogue and Interactive Systems-7

"None of the Above": Measure Uncertainty in Dialog Response Retrieval [Website][PDF]
Yulan Feng, Shikib Mehri, Maxine Eskenazi, and Tiancheng Zhao 17:00–18:00

This paper discusses the importance of uncovering uncertainty in end-to-end dialog tasks and presents our experimental results on uncertainty classification on the processed Ubuntu Dialog Corpus. We show that instead of retraining models for this specific purpose, we can capture the original retrieval model's underlying confidence concerning the best prediction using trivial additional computation.

A Generative Model for Joint Natural Language Understanding and Generation [Website][PDF]
Bo-Hsiang Tseng, Jianpeng Cheng, Yimai Fang, and David Vandyke 17:00–18:00

Natural language understanding (NLU) and natural language generation (NLG) are two fundamental and related tasks in building task-oriented dialogue systems with opposite objectives: NLU tackles the transformation from natural language to formal representations, whereas NLG does the reverse. A key to success in either task is parallel training data which is expensive to obtain at a large scale. In this work, we propose a generative model which couples NLU and NLG through a shared latent variable. This approach allows us to explore both spaces of natural language and formal representations, and facilitates information sharing through the latent space to eventually benefit NLU and NLG. Our model achieves state-of-the-art performance on two dialogue datasets with both flat and tree-structured formal representations. We also show that the model can be trained in a semi-supervised fashion by utilising unlabelled data to boost its performance.

Can You Put it All Together: Evaluating Conversational Agents' Ability to Blend Skills [Website][PDF]
Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau 17:00–18:00

Being engaging, knowledgeable, and empathetic are all desirable general qualities in a conversational agent. Previous work has introduced tasks and datasets that aim to help agents to learn those qualities in isolation and gauge how well they can express them. But rather than being specialized in one single quality, a good open-domain conversational agent should be able to seamlessly blend them all into one cohesive conversational flow. In this work, we investigate several ways to combine models trained towards isolated capabilities, ranging from simple model aggregation schemes that require minimal additional training, to various forms of multi-task training that encompass several skills at all training stages. We further propose a new dataset, BlendedSkillTalk, to analyze how these capabilities would mesh together in a natural conversation, and compare the performance of different architectures and training schemes. Our experiments show that multi-tasking over several tasks that focus on particular capabilities results in better blended conversation performance compared to models trained on a single skill, and that both unified or two-stage approaches perform well if they are constructed to avoid unwanted bias in skill selection or are fine-tuned on our new task.

Efficient Dialogue State Tracking by Selectively Overwriting Memory [Website][PDF]
Sungdong Kim, Sohee Yang, Gyuwan Kim, and Sang-Woo Lee 17:00–18:00

Recent works in dialogue state tracking (DST) focus on an open vocabulary-based setting to resolve scalability and generalization issues of the predefined ontology-based approaches. However, they are inefficient in that they predict the dialogue state at every turn from scratch. Here, we consider dialogue state as an explicit fixed-sized memory and propose a selectively overwriting mechanism for more efficient DST. This mechanism consists of two steps: (1) predicting state operation on each of the memory slots, and (2) overwriting the memory with new values, of which only a few are generated according to the predicted state operations. Our method decomposes DST into two sub-tasks and guides the decoder to focus only on one of the tasks, thus reducing the burden of the decoder. This enhances the effectiveness of training and DST performance. Our SOM-DST (Selectively Overwriting Memory for Dialogue State Tracking) model achieves state-of-the-art joint goal accuracy with 51.72% in MultiWOZ 2.0 and 53.01% in MultiWOZ 2.1 in an open vocabulary-based DST setting. In addition, we analyze the accuracy gaps between the current and the ground truth-given situations and suggest that it is a promising direction to improve state operation prediction to boost the DST performance.

Grounded Conversation Generation as Guided Traverses in Commonsense Knowledge Graphs [Website][PDF]
Houyu Zhang, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu 17:00–18:00

Human conversations naturally evolve around related concepts and hop to distant concepts. This paper presents a new conversation generation model, ConceptFlow, which leverages commonsense knowledge graphs to explicitly model conversation flows. By grounding conversations to the concept space, ConceptFlow represents the potential conversation flow as traverses in the concept space along commonsense relations. The traverse is guided by graph attentions in the concept graph, moving towards more meaningful directions in the concept space, in order to generate more semantic and informative responses. Experiments on Reddit conversations demonstrate ConceptFlow's effectiveness over previous knowledge-aware conversation models and GPT-2 based models while using 70% fewer parameters, confirming the advantage of explicit modeling conversation structures. All source codes of this work are available at https://github.com/thunlp/ConceptFlow.

Negative Training for Neural Dialogue Response Generation [Website][PDF]
Tianxing He and James Glass 17:00–18:00

Although deep learning models have brought tremendous advancements to the field of open-domain dialogue response generation, recent research results have revealed that the trained models have undesirable generation be-

haviors, such as malicious responses and generic (boring) responses. In this work, we propose a framework named "Negative Training" to minimize such behaviors. Given a trained model, the framework will first find generated samples that exhibit the undesirable behavior, and then use them to feed negative training signals for fine-tuning the model. Our experiments show that negative training can significantly reduce the hit rate of malicious responses, or discourage frequent responses and improve response diversity.

Recursive Template-based Frame Generation for Task Oriented Dialog [Website][PDF]
Rashmi Gangadharaiah and Balakrishnan Narayanaswamy 17:00–18:00

The Natural Language Understanding (NLU) component in task oriented dialog systems processes a user's request and converts it into structured information that can be consumed by downstream components such as the Dialog State Tracker (DST). This information is typically represented as a semantic frame that captures the intent and slot-labels provided by the user. We first show that such a shallow representation is insufficient for complex dialog scenarios, because it does not capture the recursive nature inherent in many domains. We propose a recursive, hierarchical frame-based representation and show how to learn it from data. We formulate the frame generation task as a template-based tree decoding task, where the decoder recursively generates a template and then fills slot values into the template. We extend local tree-based loss functions with terms that provide global supervision and show how to optimize them end-to-end. We achieve a small improvement on the widely used ATIS dataset and a much larger improvement on a more complex dataset we describe here.

Speak to your Parser: Interactive Text-to-SQL with Natural Language Feedback [Website][PDF]
Ahmed Elgohary, saghar Hosseini, and Ahmed Hassan Awadallah 17:00–18:00

We study the task of semantic parse correction with natural language feedback. Given a natural language utterance, most semantic parsing systems pose the problem as one-shot translation where the utterance is mapped to a corresponding logical form. In this paper, we investigate a more interactive scenario where humans can further interact with the system by providing free-form natural language feedback to correct the system when it generates an inaccurate interpretation of an initial utterance. We focus on natural language to SQL systems and construct, SPLASH, a dataset of utterances, incorrect SQL interpretations and the corresponding natural language feedback. We compare various reference models for the correction task and show that incorporating such a rich form of feedback can significantly improve the overall semantic parsing accuracy while retaining the flexibility of natural language interaction. While we estimated human correction accuracy is 81.5%, our best model achieves only 25.1%, which leaves a large gap for improvement in future research. SPLASH is publicly available at https://aka.ms/Splash_dataset.

[CL] The Design and Implementation of XiaoIce, an Empathetic Social Chatbot [Website][PDF]
Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum 17:00–18:00

This article describes the development of Microsoft XiaoIce, the most popular social chatbot in the world. XiaoIce is uniquely designed as an artifical intelligence companion with an emotional connection to satisfy the human need for communication, affection, and social belonging. We take into account both intelligent quotient and emotional quotient in system design, cast human–machine social chat as decision-making over Markov Decision Processes, and optimize XiaoIce for long-term user engagement, measured in expected Conversation-turns Per Session (CPS). We detail the system architecture and key components, including dialogue manager, core chat, skills, and an empathetic computing module. We show how XiaoIce dynamically recognizes human feelings and states, understands user intent, and responds to user needs throughout long conversations. Since the release in 2014, XiaoIce has communicated with over 660 million active users and succeeded in establishing long-term relationships with many of them. Analysis of large-scale online logs shows that XiaoIce has achieved an average CPS of 23, which is significantly higher than that of other chatbots and even human conversations.

USR: An Unsupervised and Reference Free Evaluation Metric for Dialog Generation [Website][PDF]
Shikib Mehri and Maxine Eskenazi 17:00–18:00

The lack of meaningful automatic evaluation metrics for dialog has impeded open-domain dialog research. Standard language generation metrics have been shown to be ineffective for evaluating dialog models. To this end, this paper presents USR, an UnSupervised and Reference-free evaluation metric for dialog. USR is a reference-free metric that trains unsupervised models to measure several desirable qualities of dialog. USR is shown to strongly correlate with human judgment on both Topical-Chat (turn-level: 0.42, system-level: 1.0) and PersonaChat (turn-level: 0.48 and system-level: 1.0). USR additionally produces interpretable measures for several desirable properties of dialog.

Zero-Shot Transfer Learning with Synthesized Data for Multi-Domain Dialogue State Tracking [Website][PDF]
Giovanni Campagna, Agata Foryciarz, Mehrad Moradshahi, and Monica Lam 17:00–18:00

Zero-shot transfer learning for multi-domain dialogue state tracking can allow us to handle new domains without incurring the high cost of data acquisition. This paper proposes new zero-short transfer learning technique for dialogue state tracking where the in-domain training data are all synthesized from an abstract dialogue model and the ontology of the domain. We show that data augmentation through synthesized data can improve the accuracy of zero-shot learning for both the TRADE model and the BERT-based SUMBT model on the MultiWOZ 2.1 dataset. We show training with only synthesized in-domain data on the SUMBT model can reach about 2/3 of the accuracy obtained with the full training dataset. We improve the zero-shot learning state of the art on average across domains by 21%.

Session 4A: Machine Learning for NLP-1

Calibrating Structured Output Predictors for Natural Language Processing [Website][PDF]
Abhyuday Jagannatha and hong yu hong 17:00–18:00

We address the problem of calibrating prediction confidence for output entities of interest in natural language processing (NLP) applications. It is important that NLP applications such as named entity recognition and question answering produce calibrated confidence scores for their predictions, especially if the applications are to be deployed in a safety-critical domain such as healthcare. However the output space of such structured prediction models are often too large to directly adapt binary or multi-class calibration methods. In this study, we propose a general calibration scheme for output entities of interest in neural network based structured prediction models. Our proposed method can be used with any binary class calibration scheme and a neural network model. Additionally, we show that our calibration method can also be used as an uncertainty-aware, entity-specific decoding step to improve the performance of the underlying model at no additional training cost or data requirements. We show that our method outperforms current calibration techniques for Named Entity Recognition, Part-of-speech tagging and Question Answering systems. We also observe an improvement in model performance from our decoding step across several tasks and benchmark datasets. Our method improves the calibration and model performance on out-of-domain test scenarios as well.

Active Imitation Learning with Noisy Guidance [Website][PDF]
Kianté Brantley, Hal Daumé III, and Amr Sharaf 17:00–18:00

Imitation learning algorithms provide state-of-the-art results on many structured prediction tasks by learning near-optimal search policies. Such algorithms assume training-time access to an expert that can provide the optimal action at any queried state; unfortunately, the number of such queries is often prohibitive, frequently rendering these approaches impractical. To combat this query complexity, we consider an active learning setting in which the learning algorithm has additional access to a much cheaper noisy heuristic that provides noisy guidance. Our algorithm, LEAQI, learns a difference classifier that predicts when the expert is likely to disagree with the heuristic, and queries the expert only when necessary. We apply LEAQI to three sequence labelling tasks, demonstrating significantly fewer queries to the expert and comparable (or better) accuracies over a passive approach.

ExpBERT: Representation Engineering with Natural Language Explanations [Website][PDF]
Shikhar Murty, Pang Wei Koh, and Percy Liang 17:00–18:00

Suppose we want to specify the inductive bias that married couples typically go on honeymoons for the task of extracting pairs of spouses from text. In this paper, we allow model developers to specify these types of inductive biases as natural language explanations. We use BERT fine-tuned on MultiNLI to "interpret" these explanations with respect to the input sentence, producing explanation-guided representations of the input. Across three relation extraction tasks, our method, ExpBERT, matches a BERT baseline but with 3–20x less labeled data and improves on the baseline by 3–10 F1 points with the same amount of labeled data.

GAN-BERT: Generative Adversarial Learning for Robust Text Classification with a Bunch of Labeled Examples [Website][PDF]
Danilo Croce, Giuseppe Castellucci, and Roberto Basili 17:00–18:00

Recent Transformer-based architectures, e.g., BERT, provide impressive results in many Natural Language Processing tasks. However, most of the adopted benchmarks are made of (sometimes hundreds of) thousands of examples. In many real scenarios, obtaining high- quality annotated data is expensive and time consuming; in contrast, unlabeled examples characterizing the target task can be, in general, easily collected. One promising method to enable semi-supervised learning has been proposed in image processing, based on Semi- Supervised Generative Adversarial Networks. In this paper, we propose GAN-BERT that ex- tends the fine-tuning of BERT-like architectures with unlabeled data in a generative adversarial setting. Experimental results show that the requirement for annotated examples can be drastically reduced (up to only 50-100 annotated examples), still obtaining good performances in several sentence classification tasks.

Generalizing Natural Language Analysis through Span-relation Representations [Website][PDF]
Zhengbao Jiang, Wei Xu, Jun Araki, and Graham Neubig 17:00–18:00

Natural language processing covers a wide variety of tasks predicting syntax, semantics, and information content, and usually each type of output is generated with specially designed architectures. In this paper, we provide the simple insight that a great variety of tasks can be represented in a single unified format consisting of labeling spans and relations between spans, thus a single task-independent model can be used across different tasks. We perform extensive experiments to test this insight on 10 disparate tasks spanning dependency parsing (syntax), semantic role labeling (semantics), relation extraction (information content), aspect based sentiment analysis (sentiment), and many others, achieving performance comparable to state-of-the-art specialized models. We further demonstrate benefits of multi-task learning, and also show that the proposed method makes it easy to analyze differences and similarities in how the model handles different tasks. Finally, we convert these datasets into a unified format to build a benchmark, which provides a holistic testbed for evaluating future models for generalized natural language analysis.

Learning to Contextually Aggregate Multi-Source Supervision for Sequence Labeling [Website][PDF]
Ouyu Lan, Xiao Huang, Bill Yuchen Lin, He Jiang, Liyuan Liu, and Xiang Ren 17:00–18:00

Sequence labeling is a fundamental task for a range of natural language processing problems. When used in practice, its performance is largely influenced by the annotation quality and quantity, and meanwhile, obtaining ground truth labels is often costly. In many cases, ground truth labels do not exist, but noisy annotations or annotations from dif-

ferent domains are accessible. In this paper, we propose a novel framework Consensus Network (ConNet) that can be trained on annotations from multiple sources (e.g., crowd annotation, cross-domain data). It learns individual representation for every source and dynamically aggregates source-specific knowledge by a context-aware attention module. Finally, it leads to a model reflecting the agreement (consensus) among multiple sources. We evaluate the proposed framework in two practical settings of multi-source learning: learning with crowd annotations and unsupervised cross-domain model adaptation. Extensive experimental results show that our model achieves significant improvements over existing methods in both settings. We also demonstrate that the method can apply to various tasks and cope with different encoders.

MixText: Linguistically-Informed Interpolation of Hidden Space for Semi-Supervised Text Classification [Website][PDF]
Jiaao Chen, Zichao Yang, and Diyi Yang 17:00–18:00

This paper presents MixText, a semi-supervised learning method for text classification, which uses our newly designed data augmentation method called TMix. TMix creates a large amount of augmented training samples by interpolating text in hidden space. Moreover, we leverage recent advances in data augmentation to guess low-entropy labels for unlabeled data, hence making them as easy to use as labeled data. By mixing labeled, unlabeled and augmented data, MixText significantly outperformed current pre-trained and fined-tuned models and other state-of-the-art semi-supervised learning methods on several text classification benchmarks. The improvement is especially prominent when supervision is extremely limited. We have publicly released our code at https://github.com/GT-SALT/MixText.

MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices [Website][PDF]
Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou 17:00–18:00

Natural Language Processing (NLP) has recently achieved great success by using huge pre-trained models with hundreds of millions of parameters. However, these models suffer from heavy model sizes and high latency such that they cannot be deployed to resource-limited mobile devices. In this paper, we propose MobileBERT for compressing and accelerating the popular BERT model. Like the original BERT, MobileBERT is task-agnostic, that is, it can be generically applied to various downstream NLP tasks via simple fine-tuning. Basically, MobileBERT is a thin version of BERT_LARGE, while equipped with bottleneck structures and a carefully designed balance between self-attentions and feed-forward networks. To train MobileBERT, we first train a specially designed teacher model, an inverted-bottleneck incorporated BERT_LARGE model. Then, we conduct knowledge transfer from this teacher to MobileBERT. Empirical studies show that MobileBERT is 4.3x smaller and 5.5x faster than BERT_BASE while achieving competitive results on well-known benchmarks. On the natural language inference tasks of GLUE, MobileBERT achieves a GLUE score of 77.7 (0.6 lower than BERT_BASE), and 62 ms latency on a Pixel 4 phone. On the SQuAD v1.1/v2.0 question answering task, MobileBERT achieves a dev F1 score of 90.0/79.2 (1.5/2.1 higher than BERT_BASE).

On Importance Sampling-Based Evaluation of Latent Language Models [Website][PDF]
Robert L Logan IV, Matt Gardner, and Sameer Singh 17:00–18:00

Language models that use additional latent structures (e.g., syntax trees, coreference chains, knowledge graph links) provide several advantages over traditional language models. However, likelihood-based evaluation of these models is often intractable as it requires marginalizing over the latent space. Existing works avoid this issue by using importance sampling. Although this approach has asymptotic guarantees, analysis is rarely conducted on the effect of decisions such as sample size and choice of proposal distribution on the reported estimates. In this paper, we carry out this analysis for three models: RNNG, EntityNLM, and KGLM. In addition, we elucidate subtle differences in how importance sampling is applied in these works that can have substantial effects on the final estimates, as well as provide theoretical results which reinforce the validity of this technique.

SMART: Robust and Efficient Fine-Tuning for Pre-trained Natural Language Models through Principled Regularized Optimization [Website][PDF]
Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao 17:00–18:00

Transfer learning has fundamentally changed the landscape of natural language processing (NLP). Many state-of-the-art models are first pre-trained on a large text corpus and then fine-tuned on downstream tasks. However, due to limited data resources from downstream tasks and the extremely high complexity of pre-trained models, aggressive fine-tuning often causes the fine-tuned model to overfit the training data of downstream tasks and fail to generalize to unseen data. To address such an issue in a principled manner, we propose a new learning framework for robust and efficient fine-tuning for pre-trained models to attain better generalization performance. The proposed framework contains two important ingredients: 1. Smoothness-inducing regularization, which effectively manages the complexity of the model; 2. Bregman proximal point optimization, which is an instance of trust-region methods and can prevent aggressive updating. Our experiments show that the proposed framework achieves new state-of-the-art performance on a number of NLP tasks including GLUE, SNLI, SciTail and ANLI. Moreover, it also outperforms the state-of-the-art T5 model, which is the largest pre-trained model containing 11 billion parameters, on GLUE.

Stolen Probability: A Structural Weakness of Neural Language Models [Website][PDF]
David Demeter, Gregory Kimmel, and Doug Downey 17:00–18:00

Neural Network Language Models (NNLMs) generate probability distributions by applying a softmax function to a distance metric formed by taking the dot product of a prediction vector with all word vectors in a high-dimensional embedding space. The dot-product distance metric forms part of the inductive bias of NNLMs. Although NNLMs optimize well with this inductive bias, we show that this results in a sub-optimal ordering of the embedding space that structurally impoverishes some words at the expense of others when assigning probability. We present numerical, theoretical and empirical analyses which show that words on the interior of the convex hull in the embedding space have their probability bounded by the probabilities of the words on the hull.

Taxonomy Construction of Unseen Domains via Graph-based Cross-Domain Knowledge Transfer [Website][PDF]
Chao Shang, Sarthak Dash, Md. Faisal Mahbub Chowdhury, Nandana Mihindukulasooriya, and Alfio Gliozzo 17:00–18:00

Extracting lexico-semantic relations as graph-structured taxonomies, also known as taxonomy construction, has been beneficial in a variety of NLP applications. Recently Graph Neural Network (GNN) has shown to be powerful in successfully tackling many tasks. However, there has been no attempt to exploit GNN to create taxonomies. In this paper, we propose Graph2Taxo, a GNN-based cross-domain transfer framework for the taxonomy construction task. Our main contribution is to learn the latent features of taxonomy construction from existing domains to guide the structure learning of an unseen domain. We also propose a novel method of directed acyclic graph (DAG) generation for taxonomy construction. Specifically, our proposed Graph2Taxo uses a noisy graph constructed from automatically extracted noisy hyponym hypernym candidate pairs, and a set of taxonomies for some known domains for training. The learned model is then used to generate taxonomy for a new unknown domain given a set of terms for that domain. Experiments on benchmark datasets from science and environment domains show that our approach attains significant improvements correspondingly over the state of the art.

To Pretrain or Not to Pretrain: Examining the Benefits of Pretrainng on Resource Rich Tasks [Website][PDF]
Sinong Wang, Madian Khabsa, and Hao Ma 17:00–18:00

Pretraining NLP models with variants of Masked Language Model (MLM) objectives has recently led to a significant improvements on many tasks. This paper examines the benefits of pretrained models as a function of the number of training samples used in the downstream task. On several text classification tasks, we show that as the number of training examples grow into the millions, the accuracy gap between finetuning BERT-based model and training vanilla LSTM from scratch narrows to within 1%. Our findings indicate that MLM-based models might reach a diminishing return point as the supervised data size increases significantly.

Why Overfitting Isn't Always Bad: Retrofitting Cross-Lingual Word Embeddings to Dictionaries [Website][PDF]
Mozhi Zhang, Yoshinari Fujinuma, Michael J. Paul, and Jordan Boyd-Graber 17:00–18:00

Cross-lingual word embeddings (CLWE) are often evaluated on bilingual lexicon induction (BLI). Recent CLWE methods use linear projections, which underfit the training dictionary, to generalize on BLI. However, underfitting can hinder generalization to other downstream tasks that rely on words from the training dictionary. We address this limitation by retrofitting CLWE to the training dictionary, which pulls training translation pairs closer in the embedding space and overfits the training dictionary. This simple post-processing step often improves accuracy on two downstream tasks, despite lowering BLI test accuracy. We also retrofit to both the training dictionary and a synthetic dictionary induced from CLWE, which sometimes generalizes even better on downstream tasks. Our results confirm the importance of fully exploiting training dictionary in downstream tasks and explains why BLI is a flawed CLWE evaluation.

XtremeDistil: Multi-stage Distillation for Massive Multilingual Models [Website][PDF]
Subhabrata Mukherjee and Ahmed Hassan Awadallah 17:00–18:00

Deep and large pre-trained language models are the state-of-the-art for various natural language processing tasks. However, the huge size of these models could be a deterrent to using them in practice. Some recent works use knowledge distillation to compress these huge models into shallow ones. In this work we study knowledge distillation with a focus on multilingual Named Entity Recognition (NER). In particular, we study several distillation strategies and propose a stage-wise optimization scheme leveraging teacher internal representations, that is agnostic of teacher architecture, and show that it outperforms strategies employed in prior works. Additionally, we investigate the role of several factors like the amount of unlabeled data, annotation resources, model architecture and inference latency to name a few. We show that our approach leads to massive compression of teacher models like mBERT by upto 35x in terms of parameters and 51x in terms of latency for batch inference while retaining 95% of its F1-score for NER over 41 languages.

Session 4A: NLP Applications-3

A Girl Has A Name: Detecting Authorship Obfuscation [Website][PDF]
Asad Mahmood, Zubair Shafiq, and Padmini Srinivasan 17:00–18:00

Authorship attribution aims to identify the author of a text based on the stylometric analysis. Authorship obfuscation, on the other hand, aims to protect against authorship attribution by modifying a text's style. In this paper, we evaluate the stealthiness of state-of-the-art authorship obfuscation methods under an adversarial threat model. An obfuscator is stealthy to the extent an adversary finds it challenging to detect whether or not a text modified by the obfuscator is obfuscated — a decision that is key to the adversary interested in authorship attribution. We show that the existing authorship obfuscation methods are not stealthy as their obfuscated texts can be identified with an average F1 score of 0.87. The reason for the lack of stealthiness is that these obfuscators degrade text smoothness, as ascertained by neural language models, in a detectable manner. Our results highlight the need to develop stealthy authorship obfuscation methods that can better protect the identity of an author seeking anonymity.

DeeBERT: Dynamic Early Exiting for Accelerating BERT Inference [Website][PDF]
Ji Xin, Raphael Tang, Jaejun Lee, Yaoliang Yu, and Jimmy Lin 17:00–18:00

Large-scale pre-trained language models such as BERT have brought significant improvements to NLP applications. However, they are also notorious for being slow in inference, which makes them difficult to deploy in real-time applications. We propose a simple but effective method, DeeBERT, to accelerate BERT inference. Our approach allows samples to exit earlier without passing through the entire model. Experiments show that DeeBERT is able to save up to ~40% inference time with minimal degradation in model quality. Further analyses show different behaviors in the BERT transformer layers and also reveal their redundancy. Our work provides new ideas to efficiently apply deep transformer-based models to downstream tasks. Code is available at https://github.com/castorini/DeeBERT.

Efficient Strategies for Hierarchical Text Classification: External Knowledge and Auxiliary Tasks [Website][PDF]
Kervy Rivas Rojas, Gina Bustamante, Arturo Oncevay, and Marco Antonio Sobrevilla Cabezudo 17:00–18:00

In hierarchical text classification, we perform a sequence of inference steps to predict the category of a document from top to bottom of a given class taxonomy. Most of the studies have focused on developing novels neural network architectures to deal with the hierarchical structure, but we prefer to look for efficient ways to strengthen a baseline model. We first define the task as a sequence-to-sequence problem. Afterwards, we propose an auxiliary synthetic task of bottom-up-classification. Then, from external dictionaries, we retrieve textual definitions for the classes of all the hierarchy's layers, and map them into the word vector space. We use the class-definition embeddings as an additional input to condition the prediction of the next layer and in an adapted beam search. Whereas the modified search did not provide large gains, the combination of the auxiliary task and the additional input of class-definitions significantly enhance the classification accuracy. With our efficient approaches, we outperform previous studies, using a drastically reduced number of parameters, in two well-known English datasets.

Investigating the effect of auxiliary objectives for the automated grading of learner English speech transcriptions [Website][PDF]
Hannah Craighead, Andrew Caines, Paula Buttery, and Helen Yannakoudakis 17:00–18:00

We address the task of automatically grading the language proficiency of spontaneous speech based on textual features from automatic speech recognition transcripts. Motivated by recent advances in multi-task learning, we develop neural networks trained in a multi-task fashion that learn to predict the proficiency level of non-native English speakers by taking advantage of inductive transfer between the main task (grading) and auxiliary prediction tasks: morpho-syntactic labeling, language modeling, and native language identification (L1). We encode the transcriptions with both bi-directional recurrent neural networks and with bi-directional representations from transformers, compare against a feature-rich baseline, and analyse performance at different proficiency levels and with transcriptions of varying error rates. Our best performance comes from a transformer encoder with L1 prediction as an auxiliary task. We discuss areas for improvement and potential applications for text-only speech scoring.

MMPE: A Multi-Modal Interface for Post-Editing Machine Translation [Website][PDF]
Nico Herbig, Tim Düwel, Santanu Pal, Kalliopi Meladaki, Mahsa Monshizadeh, Antonio Krüger, and Josef van Genabith 17:00–18:00

Current advances in machine translation (MT) increase the need for translators to switch from traditional translation to post-editing (PE) of machine-translated text, a process that saves time and reduces errors. This affects the design of translation interfaces, as the task changes from mainly generating text to correcting errors within otherwise helpful translation proposals. Since this paradigm shift offers potential for modalities other than mouse and keyboard, we present MMPE, the first prototype to combine traditional input modes with pen, touch, and speech modalities for PE of MT. The results of an evaluation with professional translators suggest that pen and touch interaction are suitable for deletion and reordering tasks, while they are of limited use for longer insertions. On the other hand, speech and multi-modal combinations of select & speech are considered suitable for replacements and insertions but offer less potential for deletion and reordering. Overall, participants were enthusiastic about the new modalities and saw them as good extensions to mouse & keyboard, but not as a complete substitute.

SPECTER: Document-level Representation Learning using Citation-informed Transformers [Website][PDF]
Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld 17:00–18:00

Representation learning is a critical ingredient for natural language processing systems. Recent Transformer language models like BERT learn powerful textual representations, but these models are targeted towards token- and sentence-level training objectives and do not leverage information on inter-document relatedness, which limits their document-level representation power. For applications on scientific documents, such as classification and recommendation, accurate embeddings of documents are a necessity. We propose SPECTER, a new method to generate document-level embedding of scientific papers based on pretraining a Transformer language model on a powerful signal of document-level relatedness: the citation graph. Unlike existing pretrained language models, Specter can be easily applied to downstream applications without task-specific fine-tuning. Additionally, to encourage further research on document-level models, we introduce SciDocs, a new evaluation benchmark consisting of seven document-level tasks ranging from citation prediction, to document classification and recommendation. We show that Specter outperforms a variety of competitive baselines on the benchmark.

Semantic Scaffolds for Pseudocode-to-Code Generation [Website][PDF]
Ruiqi Zhong, Mitchell Stern, and Dan Klein 17:00–18:00

We propose a method for program generation based on semantic scaffolds, lightweight structures representing the high-level semantic and syntactic composition of a program. By first searching over plausible scaffolds then using these as constraints for a beam search over programs, we achieve better coverage of the search space when compared with existing techniques. We apply our hierarchical search method to the SPoC dataset for pseudocode-to-code generation, in which we are given line-level natural language pseudocode annotations and aim to produce a program satisfying execution-based test cases. By using semantic scaffolds during inference, we achieve a 10% absolute improvement in top-100 accuracy over the previous state-of-the-art. Additionally, we require only 11 candidates to reach the top-3000 performance of the previous best approach when tested against unseen problems, demonstrating a substantial improvement in efficiency.

Session 4A: Question Answering-3

Contextualized Sparse Representations for Real-Time Open-Domain Question Answering [Website][PDF]
Jinhyuk Lee, Minjoon Seo, Hannaneh Hajishirzi, and Jaewoo Kang 17:00–18:00

Open-domain question answering can be formulated as a phrase retrieval problem, in which we can expect huge scalability and speed benefit but often suffer from low accuracy due to the limitation of existing phrase representation models. In this paper, we aim to improve the quality of each phrase embedding by augmenting it with a contextualized sparse representation (Sparc). Unlike previous sparse vectors that are term-frequency-based (e.g., tf-idf) or directly learned (only few thousand dimensions), we leverage rectified self-attention to indirectly learn sparse vectors in n-gram vocabulary space. By augmenting the previous phrase retrieval model (Seo et al., 2019) with Sparc, we show 4%+ improvement in CuratedTREC and SQuAD-Open. Our CuratedTREC score is even better than the best known retrieve & read model with at least 45x faster inference speed.

Dynamic Sampling Strategies for Multi-Task Reading Comprehension [Website][PDF]
Ananth Gottumukkala, Dheeru Dua, Sameer Singh, and Matt Gardner 17:00–18:00

Building general reading comprehension systems, capable of solving multiple datasets at the same time, is a recent aspirational goal in the research community. Prior work has focused on model architecture or generalization to held out datasets, and largely passed over the particulars of the multi-task learning set up. We show that a simple dynamic sampling strategy, selecting instances for training proportional to the multi-task model's current performance on a dataset relative to its single task performance, gives substantive gains over prior multi-task sampling strategies, mitigating the catastrophic forgetting that is common in multi-task learning. We also demonstrate that allowing instances of different tasks to be interleaved as much as possible between each epoch and batch has a clear benefit in multi-task performance over forcing task homogeneity at the epoch or batch level. Our final model shows greatly increased performance over the best model on ORB, a recently-released multitask reading comprehension benchmark.

Session 4A Semantics: Textual Inference and Other Areas of Semantics-1

Can We Predict New Facts with Open Knowledge Graph Embeddings? A Benchmark for Open Link Prediction [Website][PDF]
Samuel Broscheit, Kiril Gashteovski, Yanjie Wang, and Rainer Gemulla 17:00–18:00

Open Information Extraction systems extract ("subject text", "relation text", "object text") triples from raw text. Some triples are textual versions of facts, i.e., non-canonicalized mentions of entities and relations. In this paper, we investigate whether it is possible to infer new facts directly from the open knowledge graph without any canonicalization or any supervision from curated knowledge. For this purpose, we propose the open link prediction task,i.e., predicting test facts by completing ("subject text", "relation text", ?) questions. An evaluation in such a setup raises the question if a correct prediction is actually a new fact that was induced by reasoning over the open knowledge graph or if it can be trivially explained. For example, facts can appear in different paraphrased textual variants, which can lead to test leakage. To this end, we propose an evaluation protocol and a methodology for creating the open link prediction benchmark OlpBench. We performed experiments with a prototypical knowledge graph embedding model for open-link prediction. While the task is very challenging, our results suggests that it is possible to predict genuinely new facts, which can not be trivially explained.

[TACL] Decomposing Generalization: Models of Generic, Habitual and Episodic Statements [Website][PDF]
Venkata Subrahmanyan Govindarajan, Benjamin Van Durme, and Aaron Steven White 17:00–18:00

We present a novel semantic framework for modeling linguistic expressions of generalization—generic, habitual, and episodic statements—as combinations of simple, real-valued referential properties of predicates and their arguments. We use this framework to construct a dataset covering the entirety of the Universal Dependencies English Web Treebank. We use this dataset to probe the efficacy of type-level and token-level information—including hand-engineered features and static (GloVe) and contextual (ELMo) word embeddings—for predicting expressions of generalization.

INFOTABS: Inference on Tables as Semi-structured Data [Website][PDF]
Vivek Gupta, Maitrey Mehta, Pegah Nokhiz, and Vivek Srikumar 17:00–18:00

In this paper, we observe that semi-structured tabulated text is ubiquitous; understanding them requires not only comprehending the meaning of text fragments, but also implicit relationships between them. We argue that such data can prove as a testing ground for understanding how we reason about information. To study this, we introduce a new dataset called INFOTABS, comprising of human-written textual hypotheses based on premises that are tables extracted from Wikipedia info-boxes. Our analysis shows that the semi-structured, multi-domain and heterogeneous nature of the premises admits complex, multi-faceted reasoning. Experiments reveal that, while human annotators agree on the relationships between a table-hypothesis pair, several standard modeling strategies are unsuccessful at the task, suggesting that reasoning about tables can pose a difficult modeling challenge.

[TACL] Inherent Disagreements in Human Textual Inferences [Website][PDF]
Ellie Pavlick and Tom Kwiatkowski 17:00–18:00

We analyze human's disagreements about the validity of natural language inferences. We show that, very often, disagreements are not dismissible as annotation "noise", but rather persist as we collect more ratings and as we vary the amount of context provided to raters. We further show that the type of uncertainty captured by current state-of-the-art models for natural language inference is not reflective of the type of uncertainty present in human disagreements. We discuss implications of our results in relation to the recognizing textual entailment (RTE)/natural language inference (NLI) task. We argue for a refined evaluation objective which requires models to explicitly capture the full distribution of plausible human judgments.

Interactive Machine Comprehension with Information Seeking Agents [Website][PDF]
Xingdi Yuan, Jie Fu, Marc-Alexandre Côté, Yi Tay, Chris Pal, and Adam Trischler 17:00–18:00

Existing machine reading comprehension (MRC) models do not scale effectively to real-world applications like web-level information retrieval and question answering (QA). We argue that this stems from the nature of MRC datasets: most of these are static environments wherein the supporting documents and all necessary information are fully observed. In this paper, we propose a simple method that reframes existing MRC datasets as interactive, partially observable environments. Specifically, we "occlude" the majority of a document's text and add context-sensitive commands that reveal "glimpses" of the hidden text to a model. We repurpose SQuAD and NewsQA as an initial case study, and then show how the interactive corpora can be used to train a model that seeks relevant information through sequential decision making. We believe that this setting can contribute in scaling models to web-level QA scenarios.

Syntactic Data Augmentation Increases Robustness to Inference Heuristics [Website][PDF]
Junghyun Min, R. Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen 17:00–18:00

Pretrained neural models such as BERT, when fine-tuned to perform natural language inference (NLI), often show high accuracy on standard datasets, but display a surprising lack of sensitivity to word order on controlled challenge sets. We hypothesize that this issue is not primarily caused by the pretrained model's limitations, but rather by the paucity of crowdsourced NLI examples that might convey the importance of syntactic structure at the fine-tuning stage. We explore several methods to augment standard training sets with syntactically informative examples, generated by applying syntactic transformations to sentences from the MNLI corpus. The best-performing augmentation method, subject/object inversion, improved BERT's accuracy on controlled examples that diagnose sensitivity to word order from 0.28 to 0.73, without affecting performance on the MNLI test set. This improvement generalized be-

yond the particular construction used for data augmentation, suggesting that augmentation causes BERT to recruit abstract syntactic representations.

Session 4A: Speech and Multimodality-1

[TACL] Acoustic-Prosodic and Lexical Cues to Deception and Trust: Deciphering How People Detect Lies [Website] [PDF]
Xi (Leslie) Chen, Sarah Ita Levitan, Michelle Levine, Marko Mandic, and Julia Hirschberg 17:00–18:00

Humans rarely perform better than chance at lie detection. To better understand human perception of deception, we created a game framework, LieCatcher, to collect ratings of perceived deception using a large corpus of deceptive and truthful interviews. We analyzed the acoustic-prosodic and linguistic characteristics of language trusted and mistrusted by raters and compared these to characteristics of actual truthful and deceptive language to understand how perception aligns with reality. With this data we built classifiers to automatically distinguish trusted from mistrusted speech, achieving an F1 of 66.1%. We next evaluated whether the strategies raters said they used to discriminate between truthful and deceptive responses were in fact useful. Our results show that, while several prosodic and lexical features were consistently perceived as trustworthy, they were not reliable cues. Also, the strategies that judges reported using in deception detection were not helpful for the task. Our work sheds light on the nature of trusted language and provides insight into the challenging problem of human deception detection.

Improved Speech Representations with Multi-Target Autoregressive Predictive Coding [Website] [PDF]
Yu-An Chung and James Glass 17:00–18:00

Training objectives based on predictive coding have recently been shown to be very effective at learning meaningful representations from unlabeled speech. One example is Autoregressive Predictive Coding (Chung et al., 2019), which trains an autoregressive RNN to generate an unseen future frame given a context such as recent past frames. The basic hypothesis of these approaches is that hidden states that can accurately predict future frames are a useful representation for many downstream tasks. In this paper we extend this hypothesis and aim to enrich the information encoded in the hidden states by training the model to make more accurate future predictions. We propose an auxiliary objective that serves as a regularization to improve generalization of the future frame prediction task. Experimental results on phonetic classification, speech recognition, and speech translation not only support the hypothesis, but also demonstrate the effectiveness of our approach in learning representations that contain richer phonetic content.

Integrating Multimodal Information in Large Pretrained Transformers [Website] [PDF]
Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque 17:00–18:00

Recent Transformer-based contextual word representations, including BERT and XLNet, have shown state-of-the-art performance in multiple disciplines within NLP. Fine-tuning the trained contextual models on task-specific datasets has been the key to achieving superior performance downstream. While fine-tuning these pre-trained models is straightforward for lexical applications (applications with only language modality), it is not trivial for multimodal language (a growing area in NLP focused on modeling face-to-face communication). More specifically, this is due to the fact that pre-trained models don't have the necessary components to accept two extra modalities of vision and acoustic. In this paper, we proposed an attachment to BERT and XLNet called Multimodal Adaptation Gate (MAG). MAG allows BERT and XLNet to accept multimodal nonverbal data during fine-tuning. It does so by generating a shift to internal representation of BERT and XLNet; a shift that is conditioned on the visual and acoustic modalities. In our experiments, we study the commonly used CMU-MOSI and CMU-MOSEI datasets for multimodal sentiment analysis. Fine-tuning MAG-BERT and MAG-XLNet significantly boosts the sentiment analysis performance over previous baselines as well as language-only fine-tuning of BERT and XLNet. On the CMU-MOSI dataset, MAG-XLNet achieves human-level multimodal sentiment analysis performance for the first time in the NLP community.

MultiQT: Multimodal learning for real-time question tracking in speech [Website] [PDF]
Jakob D. Havtorn, Jan Latko, Joakim Edin, Lars Maaløe, Lasse Borgholt, Lorenzo Belgrano, Nicolai Jacobsen, Regitze Sdun, and Željko Agić 17:00–18:00

We address a challenging and practical task of labeling questions in speech in real time during telephone calls to emergency medical services in English, which embeds within a broader decision support system for emergency call-takers. We propose a novel multimodal approach to real-time sequence labeling in speech. Our model treats speech and its own textual representation as two separate modalities or views, as it jointly learns from streamed audio and its noisy transcription into text via automatic speech recognition. Our results show significant gains of jointly learning from the two modalities when compared to text or audio only, under adverse noise and limited volume of training data. The results generalize to medical symptoms detection where we observe a similar pattern of improvements with multimodal learning.

Multimodal and Multiresolution Speech Recognition with Transformers [Website] [PDF]
Georgios Paraskevopoulos, Srinivas Parthasarathy, Aparna Khare, and Shiva Sundaram 17:00–18:00

This paper presents an audio visual automatic speech recognition (AV-ASR) system using a Transformer-based architecture. We particularly focus on the scene context provided by the visual information, to ground the ASR. We extract representations for audio features in the encoder layers of the transformer and fuse video features using an additional crossmodal multihead attention layer. Additionally, we incorporate a multitask training criterion for multiresolution ASR, where we train the model to generate both character and subword level transcriptions. Experimental results on the How2 dataset, indicate that multiresolution training can speed up convergence by around 50% and relatively improves word error rate (WER) performance by upto 18% over subword prediction models. Further, incorporating visual information improves performance with relative gains upto 3.76% over audio only models. Our results are comparable to state-of-the-art Listen, Attend and Spell-based architectures.

Phone Features Improve Speech Translation [Website][PDF]
Elizabeth Salesky and Alan W Black 17:00–18:00

End-to-end models for speech translation (ST) more tightly couple speech recognition (ASR) and machine translation (MT) than a traditional cascade of separate ASR and MT models, with simpler model architectures and the potential for reduced error propagation. Their performance is often assumed to be superior, though in many conditions this is not yet the case. We compare cascaded and end-to-end models across high, medium, and low-resource conditions, and show that cascades remain stronger baselines. Further, we introduce two methods to incorporate phone features into ST models. We show that these features improve both architectures, closing the gap between end-to-end models and cascades, and outperforming previous academic work – by up to 9 BLEU on our low-resource setting.

Session 4A: Student Research Workshop

Media Bias, the Social Sciences, and NLP: Automating Frame Analyses to Identify Bias by Word Choice and Labeling [Website][PDF]
Felix Hamborg 17:00–18:00

Media bias can strongly impact the public perception of topics reported in the news. A difficult to detect, yet powerful form of slanted news coverage is called bias by word choice and labeling (WCL). WCL bias can occur, for example, when journalists refer to the same semantic concept by using different terms that frame the concept differently and consequently may lead to different assessments by readers, such as the terms "freedom fighters" and "terrorists," or "gun rights" and "gun control." In this research project, I aim to devise methods that identify instances of WCL bias and estimate the frames they induce, e.g., not only is "terrorists" of negative polarity but also ascribes to aggression and fear. To achieve this, I plan to research methods using natural language processing and deep learning while employing models and using analysis concepts from the social sciences, where researchers have studied media bias for decades. The first results indicate the effectiveness of this interdisciplinary research approach. My vision is to devise a system that helps news readers to become aware of the differences in media coverage caused by bias.

Exploring Interpretability in Event Extraction: Multitask Learning of a Neural Event Classifier and an Explanation Decoder [Website][PDF]
Zheng Tang, Gus Hahn-Powell, and Mihai Surdeanu 17:00–18:00

We propose an interpretable approach for event extraction that mitigates the tension between generalization and interpretability by jointly training for the two goals. Our approach uses an encoder-decoder architecture, which jointly trains a classifier for event extraction, and a rule decoder that generates syntactico-semantic rules that explain the decisions of the event classifier. We evaluate the proposed approach on three biomedical events and show that the decoder generates interpretable rules that serve as accurate explanations for the event classifier's decisions, and, importantly, that the joint training generally improves the performance of the event classifier. Lastly, we show that our approach can be used for semi-supervised learning, and that its performance improves when trained on automatically-labeled data generated by a rule-based system.

Research Replication Prediction Using Weakly Supervised Learning [Website]
Tianyi Luo, Xingyu Li, Hainan Wang, and Yang Liu 17:00–18:00

Knowing whether a published research result can be replicated or not is important. Carrying out direct replication of published research incurs high cost. It is therefore desirable to have a machine learning aided automatic prediction of a result's replicability. Such predictions can provide a confidence score for each article which can further provide guidelines for spot-checks. Since we will only have access to a small size of annotated dataset to train a machine predictor, we explore the possibility of using weakly supervised learning approaches to improve the prediction accuracy of research replication using both labelled and unlabelled datasets based on text information of research papers. Our experiments over real-world datasets show that much better prediction performance can be obtained compared to the supervised models utilizing only a small size of labelled dataset.

Crossing the Line: Where do Demographic Variables Fit into Humor Detection? [Website][PDF]
J. A. Meaney 17:00–18:00

Recent humor classification shared tasks have struggled with two issues: either the data comprises a highly constrained genre of humor which does not broadly represent humor, or the data is so indiscriminate that the inter-annotator agreement on its humor content is drastically low. These tasks typically average over all annotators' judgments, in spite of the fact that humor is a highly subjective phenomenon. We argue that demographic factors influence whether a text is perceived as humorous or not. We propose the addition of demographic information about the humor annotators in order to bin ratings more sensibly. We also suggest the addition of an 'offensive' label to distinguish between different generations, in terms of humor. This would allow for more nuanced shared tasks and could lead to better performance on downstream tasks, such as content moderation.

Demo Session 4B

Time: *17:45–18:30*

Talk to Papers: Bringing Neural Question Answering to Academic Search [Website][PDF]
Tiancheng Zhao and Kyusong Lee

We introduce Talk to Papers, which exploits the recent open-domain question answering (QA) techniques to improve the current experience of academic search. It's designed to enable researchers to use natural language queries to find precise answers and extract insights from a massive amount of academic papers. We present a large improvement over classic search engine baseline on several standard QA datasets and provide the community a collaborative data collection tool to curate the first natural language processing research QA dataset via a community effort.

BENTO: A Visual Platform for Building Clinical NLP Pipelines Based on CodaLab [Website][PDF]
Yonghao Jin, Fei Li, and Hong Yu

CodaLab is an open-source web-based platform for collaborative computational research. Although CodaLab has gained popularity in the research community, its interface has limited support for creating reusable tools that can be easily applied to new datasets and composed into pipelines. In clinical domain, natural language processing (NLP) on medical notes generally involves multiple steps, like tokenization, named entity recognition, etc. Since these steps require different tools which are usually scattered in different publications, it is not easy for researchers to use them to process their own datasets. In this paper, we present BENTO, a workflow management platform with a graphic user interface (GUI) that is built on top of CodaLab, to facilitate the process of building clinical NLP pipelines. BENTO comes with a number of clinical NLP tools that have been pre-trained using medical notes and expert annotations and can be readily used for various clinical NLP tasks. It also allows researchers and developers to create their custom tools (e.g., pre-trained NLP models) and use them in a controlled and reproducible way. In addition, the GUI interface enables researchers with limited computer background to compose tools into NLP pipelines and then apply the pipelines on their own datasets in a "what you see is what you get" (WYSIWYG) way. Although BENTO is designed for clinical NLP applications, the underlying architecture is flexible to be tailored to any other domains.

Stanza: A Python Natural Language Processing Toolkit for Many Human Languages [Website][PDF]
Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning

We introduce Stanza, an open-source Python natural language processing toolkit supporting 66 human languages. Compared to existing widely used toolkits, Stanza features a language-agnostic fully neural pipeline for text analysis, including tokenization, multi-word token expansion, lemmatization, part-of-speech and morphological feature tagging, dependency parsing, and named entity recognition. We have trained Stanza on a total of 112 datasets, including the Universal Dependencies treebanks and other multilingual corpora, and show that the same neural architecture generalizes well and achieves competitive performance on all languages tested. Additionally, Stanza includes a native Python interface to the widely used Java Stanford CoreNLP software, which further extends its functionality to cover other tasks such as coreference resolution and relation extraction. Source code, documentation, and pretrained models for 66 languages are available at https://stanfordnlp.github.io/stanza/.

Session 4B Overview – Monday, July 6, 2020 18:00–19:00

Track A *Dialogue and Interactive Systems-8* Abstracts	Grounding Conversations with Improvised Dialogues *Cho and May* [Website][PDF]	Image-Chat: Engaging Grounded Conversations *Shuster, Humeau, Bordes, and Weston* [Website][PDF]	Learning an Unreferenced Metric for Online Dialogue Evaluation *Sinha, Parthasarathi, Wang, Lowe, Hamilton, and Pineau* [Website][PDF]	Neural Generation of Dialogue Response Timings *Roddy and Harte* [Website][PDF]	The Dialogue Dodecathlon: Open-Domain Knowledge and Image Grounded Conversational Agents *Shuster, JU, Roller, Dinan, Boureau, and Weston* [Website][PDF]
Track B *Generation-7* Abstracts	Automatic Poetry Generation from Prosaic Text *Van de Cruys* [Website][PDF]	Bridging the Structural Gap Between Encoding and Decoding for Data-To-Text Generation *Zhao, Walker, and Chaturvedi* [Website][PDF]	Enabling Language Models to Fill in the Blanks *Donahue, Lee, and Liang* [Website][PDF]	GPT-too: A Language-Model-First Approach for AMR-to-Text Generation *Mager, Fernandez Astudillo, Naseem, Sultan, Lee, Florian, and Roukos* [Website][PDF]	INSET: Sentence Infilling with INter-SEntential Transformer *Huang, Zhang, Elachqar, and Cheng* [Website][PDF]
	Improving Adversarial Text Generation by Modeling the Distant Future *Zhang, Chen, Gan, Wang, Shen, Wang, Wen, and Carin* [Website][PDF]	[TACL] Leveraging Pre-trained Checkpoints for Sequence Generation Tasks *Rothe, Narayan, and Severyn* [Website][PDF]	Neural Syntactic Preordering for Controlled Paraphrase Generation *Goyal and Durrett* [Website][PDF]	Simple and Effective Retrieve-Edit-Rerank Text Generation *Hossain, Ghazvininejad, and Zettlemoyer* [Website][PDF]	
Track C *Language Grounding to Vision, Robotics and Beyond-1* Abstracts	BabyWalk: Going Farther in Vision-and-Language Navigation by Taking Baby Steps *Zhu, Hu, Chen, Deng, Jain, Ie, and Sha* [Website][PDF]	Cross-media Structured Common Space for Multimedia Event Extraction *Li, Zareian, Zeng, Whitehead, Lu, Ji, and Chang* [Website][PDF]	Learning to Segment Actions from Observation and Narration *Fried, Alayrac, Blunsom, Dyer, Clark, and Nematzadeh* [Website][PDF]	Learning to execute instructions in a Minecraft dialogue *Jayannavar, Narayan-Chen, and Hockenmaier* [Website][PDF]	MART: Memory-Augmented Recurrent Transformer for Coherent Video Paragraph Captioning *Lei, Wang, Shen, Yu, Berg, and Bansal* [Website][PDF]
	What is Learned in Visually Grounded Neural Syntax Acquisition *Kojima, Averbuch-Elor, Rush, and Artzi* [Website][PDF]				
Track D *Machine Learning for NLP-2* Abstracts	A Batch Normalized Inference Network Keeps the KL Vanishing Away *Zhu, Bi, Liu, Ma, Li, and Wu* [Website][PDF]	Contextual Embeddings: When Are They Worth It? *Arora, May, Zhang, and Ré* [Website][PDF]	[TACL] Efficient Contextual Representation Learning With Continuous Outputs *Li, Chen, Hsieh, and Chang* [Website][PDF]	Interactive Classification by Asking Informative Questions *Yu, Chen, Wang, Lei, and Artzi* [Website][PDF]	Knowledge Graph Embedding Compression *Sachan* [Website][PDF]

	Low Resource Sequence Tagging using Sentence Reconstruction *Perl, Chaudhury, and Giryes* [Website][PDF]	Masked Language Model Scoring *Salazar, Liang, Nguyen, and Kirchhoff* [Website][PDF]	Orthogonal Relation Transforms with Graph Context Modeling for Knowledge Graph Embedding *Tang, Huang, Wang, He, and Zhou* [Website][PDF]	[TACL] Perturbation Based Learning for Structured NLP Tasks with Application to Dependency Parsing *Doitch, Yazdi, Hazan, and Reichart* [Website][PDF]	Posterior Calibrated Training on Sentence Classification Tasks *Jung, Kang, Cheng, Mentch, and Schaaf* [Website][PDF]
	Posterior Control of Blackbox Generation *Li and Rush* [Website][PDF]	Pretrained Transformers Improve Out-of-Distribution Robustness *Hendrycks, Liu, Wallace, Dziedzic, Krishnan, and Song* [Website][PDF]	Robust Encodings: A Framework for Combating Adversarial Typos *Jones, Jia, Raghunathan, and Liang* [Website][PDF]	Showing Your Work Doesn't Always Work *Tang, Lee, Xin, Liu, Yu, and Lin* [Website][PDF]	Span Selection Pre-training for Question Answering *Glass, Gliozzo, Chakravarti, Ferritto, Pan, Bhargav, Garg, and Sil* [Website][PDF]
	Topological Sort for Sentence Ordering *Prabhumoye, Salakhutdinov, and Black* [Website][PDF]	Weight Poisoning Attacks on Pretrained Models *Kurita, Michel, and Neubig* [Website][PDF]	schuBERT: Optimizing Elements of BERT *Khetan and Karnin* [Website][PDF]		
Track E *Machine Translation-5* Abstracts	BPE-Dropout: Simple and Effective Subword Regularization *Provilkov, Emelianenko, and Voita* [Website][PDF]	ENGINE: Energy-Based Inference Networks for Non-Autoregressive Machine Translation *Tu, Pang, Wiseman, and Gimpel* [Website][PDF]	Leveraging Monolingual Data with Self-Supervision for Multilingual Neural Machine Translation *Siddhant, Bapna, Cao, Firat, Chen, Kudugunta, Arivazhagan, and Wu* [Website][PDF]	Multi-Domain Neural Machine Translation with Word-Level Adaptive Layer-wise Domain Mixing *Jiang, Liang, Wang, and Zhao* [Website][PDF]	On The Evaluation of Machine Translation SystemsTrained With Back-Translation *Edunov, Ott, Ranzato, and Auli* [Website][PDF]
	[CL] On the Linguistic Representational Power of Neural Machine Translation Models *Belinkov, Durrani, Dalvi, Sajjad, and Glass* [Website][PDF]	Simultaneous Translation Policies: From Fixed to Adaptive *Zheng, Liu, Zheng, Ma, Liu, and Huang* [Website][PDF]			
Track F *Lexical-2* Abstracts	Breaking Through the 80% Glass Ceiling: Raising the State of the Art in Word Sense Disambiguation by Incorporating Knowledge Graph Information *Bevilacqua and Navigli* [Website][PDF]	Glyph2Vec: Learning Chinese Out-of-Vocabulary Word Embedding from Glyphs *Chen, YU, and Lin* [Website][PDF]	[TACL] Learning Lexical Subspaces in a Distributional Vector Space *Arora, Chakraborty, and Cheung* [Website][PDF]	Moving Down the Long Tail of Word Sense Disambiguation with Gloss Informed Bi-encoders *Blevins and Zettlemoyer* [Website][PDF]	Multidirectional Associative Optimization of Function-Specific Word Representations *Gerz, Vulić, Rei, Reichart, and Korhonen* [Website][PDF]

	Predicting Degrees of Technicality in Automatic Terminology Extraction *Hätty, Schlechtweg, Dorna, and Schulte im Walde* [Website][PDF]	Verbal Multiword Expressions for Identification of Metaphor *Rohanian, Rei, Taslimipoor, and Ha* [Website][PDF]			
Track G *Student Research Workshop* Abstracts	A Geometry-Inspired Attack for Generating Natural Language Adversarial Examples *Meng and Wattenhofer* [Website]	Effectively Aligning and Filtering Parallel Corpora under Sparse Data Conditions *Steingrímsson, Loftsson, and Way* [Website][PDF]	Understanding Points of Correspondence between Sentences for Abstractive Summarization *Lebanoff, Muchovej, Dernoncourt, Kim, Wang, Chang, and Liu* [Website][PDF]	Noise-Based Augmentation Techniques for Emotion Datasets: What do we Recommend? *Jaiswal and Provost* [Website]	
Track H *Summarization-3* Abstracts	Attend to Medical Ontologies: Content Selection for Clinical Abstractive Summarization *Sotudeh Gharebagh, Goharian, and Filice* [Website][PDF]	On Faithfulness and Factuality in Abstractive Summarization *Maynez, Narayan, Bohnet, and McDonald* [Website][PDF]	Screenplay Summarization Using Latent Narrative Structure *Papalampidi, Keller, Frermann, and Lapata* [Website][PDF]	Unsupervised Opinion Summarization with Noising and Denoising *Amplayo and Lapata* [Website][PDF]	

Session 4B Details

Session 4B: Dialogue and Interactive Systems-8

Grounding Conversations with Improvised Dialogues [Website][PDF]
Hyundong Cho and Jonathan May 18:00–19:00

Effective dialogue involves grounding, the process of establishing mutual knowledge that is essential for communication between people. Modern dialogue systems are not explicitly trained to build common ground, and therefore overlook this important aspect of communication. Improvisational theater (improv) intrinsically contains a high proportion of dialogue focused on building common ground, and makes use of the yes-and principle, a strong grounding speech act, to establish coherence and an actionable objective reality. We collect a corpus of more than 26,000 yes-and turns, transcribing them from improv dialogues and extracting them from larger, but more sparsely populated movie script dialogue corpora, via a bootstrapped classifier. We fine-tune chit-chat dialogue systems with our corpus to encourage more grounded, relevant conversation and confirm these findings with human evaluations.

Image-Chat: Engaging Grounded Conversations [Website][PDF]
Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston 18:00–19:00

To achieve the long-term goal of machines being able to engage humans in conversation, our models should captivate the interest of their speaking partners. Communication grounded in images, whereby a dialogue is conducted based on a given photo, is a setup naturally appealing to humans (Hu et al., 2014). In this work we study large-scale architectures and datasets for this goal. We test a set of neural architectures using state-of-the-art image and text representations, considering various ways to fuse the components. To test such models, we collect a dataset of grounded human-human conversations, where speakers are asked to play roles given a provided emotional mood or style, as the use of such traits is also a key factor in engagingness (Guo et al., 2019). Our dataset, Image-Chat, consists of 202k dialogues over 202k images using 215 possible style traits. Automatic metrics and human evaluations of engagingness show the efficacy of our approach; in particular, we obtain state-of-the-art performance on the existing IGC task, and our best performing model is almost on par with humans on the Image-Chat test set (preferred 47.7% of the time).

Learning an Unreferenced Metric for Online Dialogue Evaluation [Website][PDF]
Koustuv Sinha, Prasanna Parthasarathi, Jasmine Wang, Ryan Lowe, William L. Hamilton, and Joelle Pineau 18:00–19:00

Evaluating the quality of a dialogue interaction between two agents is a difficult task, especially in open-domain chit-chat style dialogue. There have been recent efforts to develop automatic dialogue evaluation metrics, but most of them do not generalize to unseen datasets and/or need a human-generated reference response during inference, making it infeasible for online evaluation. Here, we propose an unreferenced automated evaluation metric that uses large pre-trained language models to extract latent representations of utterances, and leverages the temporal transitions that exist between them. We show that our model achieves higher correlation with human annotations in an online setting, while not requiring true responses for comparison during inference.

Neural Generation of Dialogue Response Timings [Website][PDF]
Matthew Roddy and Naomi Harte 18:00–19:00

The timings of spoken response offsets in human dialogue have been shown to vary based on contextual elements of the dialogue. We propose neural models that simulate the distributions of these response offsets, taking into account the response turn as well as the preceding turn. The models are designed to be integrated into the pipeline of an incremental spoken dialogue system (SDS). We evaluate our models using offline experiments as well as human listening tests. We show that human listeners consider certain response timings to be more natural based on the dialogue context. The introduction of these models into SDS pipelines could increase the perceived naturalness of interactions.

The Dialogue Dodecathlon: Open-Domain Knowledge and Image Grounded Conversational Agents [Website][PDF]
Kurt Shuster, Da JU, Stephen Roller, Emily Dinan, Y-Lan Boureau, and Jason Weston 18:00–19:00

We introduce dodecaDialogue: a set of 12 tasks that measures if a conversational agent can communicate engagingly with personality and empathy, ask questions, answer questions by utilizing knowledge resources, discuss topics and situations, and perceive and converse about images. By multi-tasking on such a broad large-scale set of data, we hope to both move towards and measure progress in producing a single unified agent that can perceive, reason and converse with humans in an open-domain setting. We show that such multi-tasking improves over a BERT pre-trained baseline, largely due to multi-tasking with very large dialogue datasets in a similar domain, and that the multi-tasking in general provides gains to both text and image-based tasks using several metrics in both the fine-tune and task transfer settings. We obtain state-of-the-art results on many of the tasks, providing a strong baseline for this challenge.

Session 4B: Generation-7

Automatic Poetry Generation from Prosaic Text [Website][PDF]
Tim Van de Cruys 18:00–19:00

In the last few years, a number of successful approaches have emerged that are able to adequately model various aspects of natural language. In particular, language models based on neural networks have improved the state of the art with regard to predictive language modeling, while topic models are successful at capturing clear-cut, semantic dimensions. In this paper, we will explore how these approaches can be adapted and combined to model the linguistic and literary aspects needed for poetry generation. The system is exclusively trained on standard, non-poetic text, and its output is constrained in order to confer a poetic character to the generated verse. The framework is applied to the generation of poems in both English and French, and is equally evaluated for both languages. Even though it only uses standard, non-poetic text as input, the system yields state of the art results for poetry generation.

Bridging the Structural Gap Between Encoding and Decoding for Data-To-Text Generation [Website][PDF]
Chao Zhao, Marilyn Walker, and Snigdha Chaturvedi 18:00–19:00

Generating sequential natural language descriptions from graph-structured data (e.g., knowledge graph) is challenging, partly because of the structural differences between the input graph and the output text. Hence, popular sequence-to-sequence models, which require serialized input, are not a natural fit for this task. Graph neural networks, on the other hand, can better encode the input graph but broaden the structural gap between the encoder and decoder, making faithful generation difficult. To narrow this gap, we propose DualEnc, a dual encoding model that can not only incorporate the graph structure, but can also cater to the linear structure of the output text. Empirical comparisons with strong single-encoder baselines demonstrate that dual encoding can significantly improve the quality of the generated text.

Enabling Language Models to Fill in the Blanks [Website][PDF]
Chris Donahue, Mina Lee, and Percy Liang 18:00–19:00

We present a simple approach for *text infilling*, the task of predicting missing spans of text at any position in a document. While infilling could enable rich functionality especially for writing assistance tools, more attention has been devoted to language modeling—a special case of infilling where text is predicted at the end of a document. In this paper, we aim to extend the capabilities of language models (LMs) to the more general task of infilling. To this end, we train (or fine tune) off-the-shelf LMs on sequences containing the concatenation of artificially-masked text and the text which was masked. We show that this approach, which we call *infilling by language modeling*, can enable LMs to infill entire sentences effectively on three different domains: short stories, scientific abstracts, and lyrics. Furthermore, we show that humans have difficulty identifying sentences infilled by our approach as machine-generated in the domain of short stories.

GPT-too: A Language-Model-First Approach for AMR-to-Text Generation [Website][PDF]
Manuel Mager, Ramón Fernandez Astudillo, Tahira Naseem, Md Arafat Sultan, Young-Suk Lee, Radu Florian, and Salim Roukos 18:00–19:00

Abstract Meaning Representations (AMRs) are broad-coverage sentence-level semantic graphs. Existing approaches to generating text from AMR have focused on training sequence-to-sequence or graph-to-sequence models on AMR annotated data only. In this paper, we propose an alternative approach that combines a strong pre-trained language model with cycle consistency-based re-scoring. Despite the simplicity of the approach, our experimental results show these models outperform all previous techniques on the English LDC2017T10 dataset, including the recent use of transformer architectures. In addition to the standard evaluation metrics, we provide human evaluation experiments that further substantiate the strength of our approach.

INSET: Sentence Infilling with INter-SEntential Transformer [Website][PDF]
Yichen Huang, Yizhe Zhang, Oussama Elachqar, and Yu Cheng 18:00–19:00

Missing sentence generation (or sentence in-filling) fosters a wide range of applications in natural language generation, such as document auto-completion and meeting note expansion. This task asks the model to generate intermediate missing sentences that can syntactically and semantically bridge the surrounding context. Solving the sentence infilling task requires techniques in natural language processing ranging from understanding to discourse-level planning to generation. In this paper, we propose a framework to decouple the challenge and address these three aspects respectively, leveraging the power of existing large-scale pre-trained models such as BERT and GPT-2. We empirically demonstrate the effectiveness of our model in learning a sentence representation for generation and further generating a missing sentence that fits the context.

Improving Adversarial Text Generation by Modeling the Distant Future [Website][PDF]
Ruiyi Zhang, Changyou Chen, Zhe Gan, Wenlin Wang, Dinghan Shen, Guoyin Wang, Zheng Wen, and Lawrence Carin 18:00–19:00

Auto-regressive text generation models usually focus on local fluency, and may cause inconsistent semantic meaning in long text generation. Further, automatically generating words with similar semantics is challenging, and hand-crafted linguistic rules are difficult to apply. We consider a text planning scheme and present a model-based imitation-learning approach to alleviate the aforementioned issues. Specifically, we propose a novel guider network to focus on the generative process over a longer horizon, which can assist next-word prediction and provide intermediate rewards for generator optimization. Extensive experiments demonstrate that the proposed method leads to improved performance.

[TACL] Leveraging Pre-trained Checkpoints for Sequence Generation Tasks [Website][PDF]
Sascha Rothe, Shashi Narayan, and Aliaksei Severyn 18:00–19:00

Unsupervised pre-training of large neural models has recently revolutionized Natural Language Processing. By warm-starting from the publicly released checkpoints, NLP practitioners have pushed the state-of-the-art on multiple benchmarks while saving significant amounts of compute time. So far the focus has been mainly on the Natural Language Understanding tasks. In this paper, we demonstrate the efficacy of pre-trained checkpoints for Sequence Generation. We developed a Transformer-based sequence-to-sequence model that is compatible with publicly available pre-trained BERT, GPT-2 and RoBERTa checkpoints and conducted an extensive empirical study on the utility of initializing our model, both encoder and decoder, with these checkpoints. Our models result in new state-of-the-art results on Machine Translation, Text Summarization, Sentence Splitting, and Sentence Fusion.

Neural Syntactic Preordering for Controlled Paraphrase Generation [Website][PDF]
Tanya Goyal and Greg Durrett 18:00–19:00

Paraphrasing natural language sentences is a multifaceted process: it might involve replacing individual words or short phrases, local rearrangement of content, or high-level restructuring like topicalization or passivization. Past approaches struggle to cover this space of paraphrase possibilities in an interpretable manner. Our work, inspired by pre-ordering literature in machine translation, uses syntactic transformations to softly "reorder" the source sentence and guide our neural paraphrasing model. First, given an input sentence, we derive a set of feasible syntactic rearrangements using an encoder-decoder model. This model operates over a partially lexical, partially syntactic view of the sentence and can reorder big chunks. Next, we use each proposed rearrangement to produce a sequence of position embeddings, which encourages our final encoder-decoder paraphrase model to attend to the source words in a particular order. Our evaluation, both automatic and human, shows that the proposed system retains the quality of the baseline approaches while giving a substantial increase in the diversity of the generated paraphrases.

Simple and Effective Retrieve-Edit-Rerank Text Generation [Website][PDF]
Nabil Hossain, Marjan Ghazvininejad, and Luke Zettlemoyer 18:00–19:00

Retrieve-and-edit seq2seq methods typically retrieve an output from the training set and learn a model to edit it to produce the final output. We propose to extend this framework with a simple and effective post-generation ranking approach. Our framework (i) retrieves several potentially relevant outputs for each input, (ii) edits each candidate independently, and (iii) re-ranks the edited candidates to select the final output. We use a standard editing model with simple task-specific re-ranking approaches, and we show empirically that this approach outperforms existing, significantly more complex methodologies. Experiments on two machine translation (MT) datasets show new state-of-art results. We also achieve near state-of-art performance on the Gigaword summarization dataset, where our analyses show that there is significant room for performance improvement with better candidate output selection in future work.

Session 4B: Language Grounding to Vision, Robotics and Beyond-1

BabyWalk: Going Farther in Vision-and-Language Navigation by Taking Baby Steps [Website][PDF]
Wang Zhu, Hexiang Hu, Jiacheng Chen, Zhiwei Deng, Vihan Jain, Eugene Ie, and Fei Sha 18:00–19:00

Learning to follow instructions is of fundamental importance to autonomous agents for vision-and-language navigation (VLN). In this paper, we study how an agent can navigate long paths when learning from a corpus that consists of shorter ones. We show that existing state-of-the-art agents do not generalize well. To this end, we propose BabyWalk, a new VLN agent that is learned to navigate by decomposing long instructions into shorter ones (BabySteps) and completing them sequentially. A special design memory buffer is used by the agent to turn its past experiences into contexts for future steps. The learning process is composed of two phases. In the first phase, the agent uses imitation learning from demonstration to accomplish BabySteps. In the second phase, the agent uses curriculum-based reinforcement learning to maximize rewards on navigation tasks with increasingly longer instructions. We create two new benchmark datasets (of long navigation tasks) and use them in conjunction with existing ones to examine BabyWalk's generalization ability. Empirical results show that BabyWalk achieves state-of-the-art results on several metrics, in particular, is able to follow long instructions better. The codes and the datasets are released on our project page: https://github.com/Sha-Lab/babywalk.

Cross-media Structured Common Space for Multimedia Event Extraction [Website][PDF]
Manling Li, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, and Shih-Fu Chang 18:00–19:00

We introduce a new task, MultiMedia Event Extraction, which aims to extract events and their arguments from multimedia documents. We develop the first benchmark and collect a dataset of 245 multimedia news articles with extensively annotated events and arguments. We propose a novel method, Weakly Aligned Structured Embedding (WASE), that encodes structured representations of semantic information from textual and visual data into a common embedding space. The structures are aligned across modalities by employing a weakly supervised training strategy, which enables exploiting available resources without explicit cross-media annotation. Compared to uni-modal state-of-the-art methods, our approach achieves 4.0% and 9.8% absolute F-score gains on text event argument role labeling and visual event extraction. Compared to state-of-the-art multimedia unstructured representations, we achieve 8.3% and 5.0% absolute F-score gains on multimedia event extraction and argument role labeling, respectively. By utilizing images, we extract 21.4% more event mentions than traditional text-only methods.

Learning to Segment Actions from Observation and Narration [Website][PDF]
Daniel Fried, Jean-Baptiste Alayrac, Phil Blunsom, Chris Dyer, Stephen Clark, and Aida Nematzadeh 18:00–19:00

We apply a generative segmental model of task structure, guided by narration, to action segmentation in video. We focus on unsupervised and weakly-supervised settings where no action labels are known during training. Despite its simplicity, our model performs competitively with previous work on a dataset of naturalistic instructional videos. Our model allows us to vary the sources of supervision used in training, and we find that both task structure and narrative language provide large benefits in segmentation quality.

Learning to execute instructions in a Minecraft dialogue [Website][PDF]
Prashant Jayannavar, Anjali Narayan-Chen, and Julia Hockenmaier 18:00–19:00

The Minecraft Collaborative Building Task is a two-player game in which an Architect (A) instructs a Builder (B) to construct a target structure in a simulated Blocks World Environment. We define the subtask of predicting correct action sequences (block placements and removals) in a given game context, and show that capturing B's past actions as well as B's perspective leads to a significant improvement in performance on this challenging language understanding problem.

MART: Memory-Augmented Recurrent Transformer for Coherent Video Paragraph Captioning [Website][PDF]
Jie Lei, Liwei Wang, Yelong Shen, Dong Yu, Tamara Berg, and Mohit Bansal 18:00–19:00

Generating multi-sentence descriptions for videos is one of the most challenging captioning tasks due to its high requirements for not only visual relevance but also discourse-based coherence across the sentences in the paragraph. Towards this goal, we propose a new approach called Memory-Augmented Recurrent Transformer (MART), which uses a memory module to augment the transformer architecture. The memory module generates a highly summarized memory state from the video segments and the sentence history so as to help better prediction of the next sentence (w.r.t. coreference and repetition aspects), thus encouraging coherent paragraph generation. Extensive experiments, human evaluations, and qualitative analyses on two popular datasets ActivityNet Captions and YouCookII show that MART generates more coherent and less repetitive paragraph captions than baseline methods, while maintaining relevance to the input video events.

What is Learned in Visually Grounded Neural Syntax Acquisition [Website][PDF]
Noriyuki Kojima, Hadar Averbuch-Elor, Alexander Rush, and Yoav Artzi 18:00–19:00

Visual features are a promising signal for learning bootstrap textual models. However, blackbox learning models make it difficult to isolate the specific contribution of visual components. In this analysis, we consider the case study of the Visually Grounded Neural Syntax Learner (Shi et al., 2019), a recent approach for learning syntax from a visual training signal. By constructing simplified versions of the model, we isolate the core factors that yield the model's strong performance. Contrary to what the model might be capable of learning, we find significantly less expressive versions produce similar predictions and perform just as well, or even better. We also find that a simple lexical signal

of noun concreteness plays the main role in the model's predictions as opposed to more complex syntactic reasoning.

Session 4B: Machine Learning for NLP-2

A Batch Normalized Inference Network Keeps the KL Vanishing Away [Website][PDF]
Qile Zhu, Wei Bi, Xiaojiang Liu, Xiyao Ma, Xiaolin Li, and Dapeng Wu 18:00–19:00

Variational Autoencoder (VAE) is widely used as a generative model to approximate a model's posterior on latent variables by combining the amortized variational inference and deep neural networks. However, when paired with strong autoregressive decoders, VAE often converges to a degenerated local optimum known as "posterior collapse". Previous approaches consider the Kullback—Leibler divergence (KL) individual for each datapoint. We propose to let the KL follow a distribution across the whole dataset, and analyze that it is sufficient to prevent posterior collapse by keeping the expectation of the KL's distribution positive. Then we propose Batch Normalized-VAE (BN-VAE), a simple but effective approach to set a lower bound of the expectation by regularizing the distribution of the approximate posterior's parameters. Without introducing any new model component or modifying the objective, our approach can avoid the posterior collapse effectively and efficiently. We further show that the proposed BN-VAE can be extended to conditional VAE (CVAE). Empirically, our approach surpasses strong autoregressive baselines on language modeling, text classification and dialogue generation, and rivals more complex approaches while keeping almost the same training time as VAE.

Contextual Embeddings: When Are They Worth It? [Website][PDF]
Simran Arora, Avner May, Jian Zhang, and Christopher Ré 18:00–19:00

We study the settings for which deep contextual embeddings (e.g., BERT) give large improvements in performance relative to classic pretrained embeddings (e.g., GloVe), and an even simpler baseline—random word embeddings—focusing on the impact of the training set size and the linguistic properties of the task. Surprisingly, we find that both of these simpler baselines can match contextual embeddings on industry-scale data, and often perform within 5 to 10% accuracy (absolute) on benchmark tasks. Furthermore, we identify properties of data for which contextual embeddings give particularly large gains: language containing complex structure, ambiguous word usage, and words unseen in training.

[TACL] Efficient Contextual Representation Learning With Continuous Outputs [Website][PDF]
Liunian Harold Li, Patrick H. Chen, Cho-Jui Hsieh, and Kai-Wei Chang 18:00–19:00

Contextual representation models have achieved great success in improving various downstream natural language processing tasks. However, these language-model-based encoders are difficult to train due to their large parameter size and high computational complexity. By carefully examining the training procedure, we observe that the softmax layer, which predicts a distribution of the target word, often induces significant overhead, especially when the vocabulary size is large. Therefore, we revisit the design of the output layer and consider directly predicting the pre-trained embedding of the target word for a given context. When applied to ELMo, the proposed approach achieves a 4 times speedup and eliminates 80% trainable parameters while achieving competitive performance on downstream tasks. Further analysis shows that the approach maintains the speed advantage under various settings, even when the sentence encoder is scaled up.

Interactive Classification by Asking Informative Questions [Website][PDF]
Lili Yu, Howard Chen, Sida I. Wang, Tao Lei, and Yoav Artzi 18:00–19:00

We study the potential for interaction in natural language classification. We add a limited form of interaction for intent classification, where users provide an initial query using natural language, and the system asks for additional information using binary or multi-choice questions. At each turn, our system decides between asking the most informative question or making the final classification pre-diction. The simplicity of the model allows for bootstrapping of the system without interaction data, instead relying on simple crowd-sourcing tasks. We evaluate our approach on two domains, showing the benefit of interaction and the advantage of learning to balance between asking additional questions and making the final prediction.

Knowledge Graph Embedding Compression [Website][PDF]
Mrinmaya Sachan 18:00–19:00

Knowledge graph (KG) representation learning techniques that learn continuous embeddings of entities and relations in the KG have become popular in many AI applications. With a large KG, the embeddings consume a large amount of storage and memory. This is problematic and prohibits the deployment of these techniques in many real world settings. Thus, we propose an approach that compresses the KG embedding layer by representing each entity in the KG as a vector of discrete codes and then composes the embeddings from these codes. The approach can be trained end-to-end with simple modifications to any existing KG embedding technique. We evaluate the approach on various standard KG embedding evaluations and show that it achieves 50-1000x compression of embeddings with a minor loss in performance. The compressed embeddings also retain the ability to perform various reasoning tasks such as KG inference.

Low Resource Sequence Tagging using Sentence Reconstruction [Website][PDF]
Tal Perl, Sriram Chaudhury, and Raja Giryes 18:00–19:00

This work revisits the task of training sequence tagging models with limited resources using transfer learning. We investigate several proposed approaches introduced in recent works and suggest a new loss that relies on sentence reconstruction from normalized embeddings. Specifically, our method demonstrates how by adding a decoding layer for sentence reconstruction, we can improve the performance of various baselines. We show improved results on the CoNLL02 NER and UD 1.2 POS datasets and demonstrate the power of the method for transfer learning with low-resources achieving 0.6 F1 score in Dutch using only one sample from it.

Masked Language Model Scoring [Website][PDF]
Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff 18:00–19:00

Pretrained masked language models (MLMs) require finetuning for most NLP tasks. Instead, we evaluate MLMs out of the box via their pseudo-log-likelihood scores (PLLs), which are computed by masking tokens one by one. We show that PLLs outperform scores from autoregressive language models like GPT-2 in a variety of tasks. By rescoring ASR and NMT hypotheses, RoBERTa reduces an end-to-end LibriSpeech model's WER by 30% relative and adds up to +1.7 BLEU on state-of-the-art baselines for low-resource translation pairs, with further gains from domain adaptation. We attribute this success to PLL's unsupervised expression of linguistic acceptability without a left-to-right bias, greatly improving on scores from GPT-2 (+10 points on island effects, NPI licensing in BLiMP). One can finetune MLMs to give scores without masking, enabling computation in a single inference pass. In all, PLLs and their associated pseudo-perplexities (PPPLs) enable plug-and-play use of the growing number of pretrained MLMs; e.g., we use a single cross-lingual model to rescore translations in multiple languages. We release our library for language model scoring at https://github.com/awslabs/mlm-scoring.

Orthogonal Relation Transforms with Graph Context Modeling for Knowledge Graph Embedding [Website][PDF]
Yun Tang, Jing Huang, Guangtao Wang, Xiaodong He, and Bowen Zhou 18:00–19:00

Distance-based knowledge graph embeddings have shown substantial improvement on the knowledge graph link prediction task, from TransE to the latest state-of-the-art RotatE. However, complex relations such as N-to-1, 1-to-N and N-to-N still remain challenging to predict. In this work, we propose a novel distance-based approach for knowledge graph link prediction. First, we extend the RotatE from 2D complex domain to high dimensional space with orthogonal transforms to model relations. The orthogonal transform embedding for relations keeps the capability for modeling symmetric/anti-symmetric, inverse and compositional relations while achieves better modeling capacity. Second, the graph context is integrated into distance scoring functions directly. Specifically, graph context is explicitly modeled via two directed context representations. Each node embedding in knowledge graph is augmented with two context representations, which are computed from the neighboring outgoing and incoming nodes/edges respectively. The proposed approach improves prediction accuracy on the difficult N-to-1, 1-to-N and N-to-N cases. Our experimental results show that it achieves state-of-the-art results on two common benchmarks FB15k-237 and WNRR-18, especially on FB15k-237 which has many high in-degree nodes.

[TACL] Perturbation Based Learning for Structured NLP Tasks with Application to Dependency Parsing [Website][PDF]
Amichay Doitch, Ram Yazdi, Tamir Hazan, and Roi Reichart 18:00–19:00

The best solution of structured prediction models in NLP is often inaccurate due to limited expressive power of the model or to non-exact parameter estimation. One way to mitigate this problem is sampling candidate solutions from the model's solution space, reasoning that effective exploration of this space should yield high quality solutions. Unfortunately, sampling is often computationally hard and many works hence back-off to sub-optimal strategies such as extraction of the best scoring solutions of the model, which are not as diverse as sampled solutions. In this paper we propose a perturbation-based approach where sampling from a probabilistic model is computationally efficient. We present a learning algorithm for the variance of the perturbations, and empirically demonstrate its importance. Moreover, while finding the argmax in our model is intractable, we propose an efficient and effective approximation. We apply our framework to cross-lingual dependency parsing across 72 corpora from 42 languages and to lightly supervised dependency parsing across 13 corpora from 12 languages and demonstrate strong results in terms of both the quality of the entire solution list and of the final solution.

Posterior Calibrated Training on Sentence Classification Tasks [Website][PDF]
Taehee Jung, Dongyeop Kang, Hua Cheng, Lucas Mentch, and Thomas Schaaf 18:00–19:00

Most classification models work by first predicting a posterior probability distribution over all classes and then selecting that class with the largest estimated probability. In many settings however, the quality of posterior probability itself (e.g., 65% chance having diabetes), gives more reliable information than the final predicted class alone. When these methods are shown to be poorly calibrated, most fixes to date have relied on posterior calibration, which rescales the predicted probabilities but often has little impact on final classifications. Here we propose an end-to-end training procedure called posterior calibrated (PosCal) training that directly optimizes the objective while minimizing the difference between the predicted and empirical posterior probabilities.We show that PosCal not only helps reduce the calibration error but also improve task performance by penalizing drops in performance of both objectives. Our PosCal achieves about 2.5% of task performance gain and 16.1% of calibration error reduction on GLUE (Wang et al., 2018) compared to the baseline. We achieved the comparable task performance with 13.2% calibration error reduction on xSLUE (Kang and Hovy, 2019), but not outperforming the two-stage calibration baseline. PosCal training can be easily extendable to any types of classification tasks as a form of regularization term. Also, PosCal has the advantage that it incrementally tracks needed statistics for the calibration objective during the training process, making efficient use of large training sets.

Posterior Control of Blackbox Generation [Website][PDF]
Xiang Lisa Li and Alexander Rush 18:00–19:00

Text generation often requires high-precision output that obeys task-specific rules. This fine-grained control is difficult to enforce with off-the-shelf deep learning models. In this work, we consider augmenting neural generation models with discrete control states learned through a structured latent-variable approach. Under this formulation, task-specific knowledge can be encoded through a range of rich, posterior constraints that are effectively trained into the model. This approach allows users to ground internal model decisions based on prior knowledge, without sacrificing the representational power of neural generative models. Experiments consider applications of this approach

for text generation. We find that this method improves over standard benchmarks, while also providing fine-grained control.

Pretrained Transformers Improve Out-of-Distribution Robustness [Website][PDF]
Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song 18:00–19:00

Although pretrained Transformers such as BERT achieve high accuracy on in-distribution examples, do they generalize to new distributions? We systematically measure out-of-distribution (OOD) generalization for seven NLP datasets by constructing a new robustness benchmark with realistic distribution shifts. We measure the generalization of previous models including bag-of-words models, ConvNets, and LSTMs, and we show that pretrained Transformers' performance declines are substantially smaller. Pretrained transformers are also more effective at detecting anomalous or OOD examples, while many previous models are frequently worse than chance. We examine which factors affect robustness, finding that larger models are not necessarily more robust, distillation can be harmful, and more diverse pretraining data can enhance robustness. Finally, we show where future work can improve OOD robustness.

Robust Encodings: A Framework for Combating Adversarial Typos [Website][PDF]
Erik Jones, Robin Jia, Aditi Raghunathan, and Percy Liang 18:00–19:00

Despite excellent performance on many tasks, NLP systems are easily fooled by small adversarial perturbations of inputs. Existing procedures to defend against such perturbations are either (i) heuristic in nature and susceptible to stronger attacks or (ii) provide guaranteed robustness to worst-case attacks, but are incompatible with state-of-the-art models like BERT. In this work, we introduce robust encodings (RobEn): a simple framework that confers guaranteed robustness, without making compromises on model architecture. The core component of RobEn is an encoding function, which maps sentences to a smaller, discrete space of encodings. Systems using these encodings as a bottleneck confer guaranteed robustness with standard training, and the same encodings can be used across multiple tasks. We identify two desiderata to construct robust encoding functions: perturbations of a sentence should map to a small set of encodings (stability), and models using encodings should still perform well (fidelity). We instantiate RobEn to defend against a large family of adversarial typos. Across six tasks from GLUE, our instantiation of RobEn paired with BERT achieves an average robust accuracy of 71.3% against all adversarial typos in the family considered, while previous work using a typo-corrector achieves only 35.3% accuracy against a simple greedy attack.

Showing Your Work Doesn't Always Work [Website][PDF]
Raphael Tang, Jaejun Lee, Ji Xin, Xinyu Liu, Yaoliang Yu, and Jimmy Lin 18:00–19:00

In natural language processing, a recently popular line of work explores how to best report the experimental results of neural networks. One exemplar publication, titled "Show Your Work: Improved Reporting of Experimental Results" (Dodge et al., 2019), advocates for reporting the expected validation effectiveness of the best-tuned model, with respect to the computational budget. In the present work, we critically examine this paper. As far as statistical generalizability is concerned, we find unspoken pitfalls and caveats with this approach. We analytically show that their estimator is biased and uses error-prone assumptions. We find that the estimator favors negative errors and yields poor bootstrapped confidence intervals. We derive an unbiased alternative and bolster our claims with empirical evidence from statistical simulation. Our codebase is at https://github.com/castorini/meanmax.

Span Selection Pre-training for Question Answering [Website][PDF]
Michael Glass, Alfio Gliozzo, Rishav Chakravarti, Anthony Ferritto, Lin Pan, G P Shrivatsa Bhargav, Dinesh Garg, and Avi Sil 18:00–19:00

BERT (Bidirectional Encoder Representations from Transformers) and related pre-trained Transformers have provided large gains across many language understanding tasks, achieving a new state-of-the-art (SOTA). BERT is pre-trained on two auxiliary tasks: Masked Language Model and Next Sentence Prediction. In this paper we introduce a new pre-training task inspired by reading comprehension to better align the pre-training from memorization to understanding. Span Selection PreTraining (SSPT) poses cloze-like training instances, but rather than draw the answer from the model's parameters, it is selected from a relevant passage. We find significant and consistent improvements over both BERT-BASE and BERT-LARGE on multiple Machine Reading Comprehension (MRC) datasets. Specifically, our proposed model has strong empirical evidence as it obtains SOTA results on Natural Questions, a new benchmark MRC dataset, outperforming BERT-LARGE by 3 F1 points on short answer prediction. We also show significant impact in HotpotQA, improving answer prediction F1 by 4 points and supporting fact prediction F1 by 1 point and outperforming the previous best system. Moreover, we show that our pre-training approach is particularly effective when training data is limited, improving the learning curve by a large amount.

Topological Sort for Sentence Ordering [Website][PDF]
Shrimai Prabhumoye, Ruslan Salakhutdinov, and Alan W Black 18:00–19:00

Sentence ordering is the task of arranging the sentences of a given text in the correct order. Recent work using deep neural networks for this task has framed it as a sequence prediction problem. In this paper, we propose a new framing of this task as a constraint solving problem and introduce a new technique to solve it. Additionally, we propose a human evaluation for this task. The results on both automatic and human metrics across four different datasets show that this new technique is better at capturing coherence in documents.

Weight Poisoning Attacks on Pretrained Models [Website][PDF]
Keita Kurita, Paul Michel, and Graham Neubig 18:00–19:00

Recently, NLP has seen a surge in the usage of large pre-trained models. Users download weights of models pre-trained on large datasets, then fine-tune the weights on a task of their choice. This raises the question of whether downloading untrusted pre-trained weights can pose a security threat. In this paper, we show that it is possible to

construct "weight poisoning" attacks where pre-trained weights are injected with vulnerabilities that expose "backdoors" after fine-tuning, enabling the attacker to manipulate the model prediction simply by injecting an arbitrary keyword. We show that by applying a regularization method which we call RIPPLe and an initialization procedure we call Embedding Surgery, such attacks are possible even with limited knowledge of the dataset and fine-tuning procedure. Our experiments on sentiment classification, toxicity detection, and spam detection show that this attack is widely applicable and poses a serious threat. Finally, we outline practical defenses against such attacks.[1]

schuBERT: Optimizing Elements of BERT [Website] [PDF]

Ashish Khetan and Zohar Karnin 18:00–19:00

Transformers have gradually become a key component for many state-of-the-art natural language representation models. A recent Transformer based model- BERTachieved state-of-the-art results on various natural language processing tasks, including GLUE, SQuAD v1.1, and SQuAD v2.0. This model however is computationally prohibitive and has a huge number of parameters. In this work we revisit the architecture choices of BERT in efforts to obtain a lighter model. We focus on reducing the number of parameters yet our methods can be applied towards other objectives such FLOPs or latency. We show that much efficient light BERT models can be obtained by reducing algorithmically chosen correct architecture design dimensions rather than reducing the number of Transformer encoder layers. In particular, our schuBERT gives 6.6% higher average accuracy on GLUE and SQuAD datasets as compared to BERT with three encoder layers while having the same number of parameters.

[1] Our code will be made publicly available on publication.

Session 4B: Machine Translation-5

BPE-Dropout: Simple and Effective Subword Regularization [Website][PDF]
Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita 18:00–19:00

Subword segmentation is widely used to address the open vocabulary problem in machine translation. The dominant approach to subword segmentation is Byte Pair Encoding (BPE), which keeps the most frequent words intact while splitting the rare ones into multiple tokens. While multiple segmentations are possible even with the same vocabulary, BPE splits words into unique sequences; this may prevent a model from better learning the compositionality of words and being robust to segmentation errors. So far, the only way to overcome this BPE imperfection, its deterministic nature, was to create another subword segmentation algorithm (Kudo, 2018). In contrast, we show that BPE itself incorporates the ability to produce multiple segmentations of the same word. We introduce BPE-dropout - simple and effective subword regularization method based on and compatible with conventional BPE. It stochastically corrupts the segmentation procedure of BPE, which leads to producing multiple segmentations within the same fixed BPE framework. Using BPE-dropout during training and the standard BPE during inference improves translation quality up to 2.3 BLEU compared to BPE and up to 0.9 BLEU compared to the previous subword regularization.

ENGINE: Energy-Based Inference Networks for Non-Autoregressive Machine Translation [Website][PDF]
Lifu Tu, Richard Yuanzhe Pang, Sam Wiseman, and Kevin Gimpel 18:00–19:00

We propose to train a non-autoregressive machine translation model to minimize the energy defined by a pretrained autoregressive model. In particular, we view our non-autoregressive translation system as an inference network (Tu and Gimpel, 2018) trained to minimize the autoregressive teacher energy. This contrasts with the popular approach of training a non-autoregressive model on a distilled corpus consisting of the beam-searched outputs of such a teacher model. Our approach, which we call ENGINE (ENerGy-based Inference NEtworks), achieves state-of-the-art non-autoregressive results on the IWSLT 2014 DE-EN and WMT 2016 RO-EN datasets, approaching the performance of autoregressive models.

Leveraging Monolingual Data with Self-Supervision for Multilingual Neural Machine Translation [Website][PDF]
Aditya Siddhant, Ankur Bapna, Yuan Cao, Orhan Firat, Mia Chen, Sneha Kudugunta, Naveen Arivazhagan, and Yonghui Wu 18:00–19:00

Over the last few years two promising research directions in low-resource neural machine translation (NMT) have emerged. The first focuses on utilizing high-resource languages to improve the quality of low-resource languages via multilingual NMT. The second direction employs monolingual data with self-supervision to pre-train translation models, followed by fine-tuning on small amounts of supervised data. In this work, we join these two lines of research and demonstrate the efficacy of monolingual data with self-supervision in multilingual NMT. We offer three major results: (i) Using monolingual data significantly boosts the translation quality of low-resource languages in multilingual models. (ii) Self-supervision improves zero-shot translation quality in multilingual models. (iii) Leveraging monolingual data with self-supervision provides a viable path towards adding new languages to multilingual models, getting up to 33 BLEU on ro-en translation without any parallel data or back-translation.

Multi-Domain Neural Machine Translation with Word-Level Adaptive Layer-wise Domain Mixing [Website][PDF]
Haoming Jiang, Chen Liang, Chong Wang, and Tuo Zhao 18:00–19:00

Many multi-domain neural machine translation (NMT) models achieve knowledge transfer by enforcing one encoder to learn shared embedding across domains. However, this design lacks adaptation to individual domains. To overcome this limitation, we propose a novel multi-domain NMT model using individual modules for each domain, on which we apply word-level, adaptive and layer-wise domain mixing. We first observe that words in a sentence are often related to multiple domains. Hence, we assume each word has a domain proportion, which indicates its domain preference. Then word representations are obtained by mixing their embedding in individual domains based on their domain proportions. We show this can be achieved by carefully designing multi-head dot-product attention modules for different domains, and eventually taking weighted averages of their parameters by word-level layer-wise domain proportions. Through this, we can achieve effective domain knowledge sharing and capture fine-grained domain-specific knowledge as well. Our experiments show that our proposed model outperforms existing ones in several NMT tasks.

On The Evaluation of Machine Translation SystemsTrained With Back-Translation [Website][PDF]
Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, and Michael Auli 18:00–19:00

Back-translation is a widely used data augmentation technique which leverages target monolingual data. However, its effectiveness has been challenged since automatic metrics such as BLEU only show significant improvements for test examples where the source itself is a translation, or translationese. This is believed to be due to translationese inputs better matching the back-translated training data. In this work, we show that this conjecture is not empirically supported and that back-translation improves translation quality of both naturally occurring text as well as translationese according to professional human translators. We provide empirical evidence to support the view that back-translation is preferred by humans because it produces more fluent outputs. BLEU cannot capture human preferences because references are translationese when source sentences are natural text. We recommend complementing BLEU with a language model score to measure fluency.

[CL] On the Linguistic Representational Power of Neural Machine Translation Models [Website][PDF]
Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass 18:00–19:00

Despite the recent success of deep neural networks in natural language processing and other spheres of artificial intelligence, their interpretability remains a challenge. We analyze the representations learned by neural machine translation (NMT) models at various levels of granularity and evaluate their quality through relevant extrinsic properties. In particular, we seek answers to the following questions: (i) How accurately is word structure captured within the learned representations, which is an important aspect in translating morphologically rich languages? (ii) Do the representations capture long-range dependencies, and effectively handle syntactically divergent languages? (iii) Do the representations capture lexical semantics? We conduct a thorough investigation along several parameters: (i) Which layers in the architecture capture each of these linguistic phenomena; (ii) How does the choice of translation unit (word, character, or subword unit) impact the linguistic properties captured by the underlying representations? (iii) Do the encoder and decoder learn differently and independently? (iv) Do the representations learned by multilingual NMT models capture the same amount of linguistic information as their bilingual counterparts? Our data-driven, quantitative evaluation illuminates important aspects in NMT models and their ability to capture various linguistic phenomena. We show that deep NMT models trained in an end-to-end fashion, without being provided any direct supervision during the training process, learn a non-trivial amount of linguistic information. Notable findings include the following observations: (i) Word morphology and part-of-speech information are captured at the lower layers of the model; (ii) In contrast, lexical semantics or non-local syntactic and semantic dependencies are better represented at the higher layers of the model; (iii) Representations learned using characters are more informed about word-morphology compared to those learned using subword units; and (iv) Representations learned by multilingual models are richer compared to bilingual models.

Simultaneous Translation Policies: From Fixed to Adaptive [Website][PDF]
Baigong Zheng, Kaibo Liu, Renjie Zheng, Mingbo Ma, Hairong Liu, and Liang Huang 18:00–19:00

Adaptive policies are better than fixed policies for simultaneous translation, since they can flexibly balance the trade-off between translation quality and latency based on the current context information. But previous methods on obtaining adaptive policies either rely on complicated training process, or underperform simple fixed policies. We design an algorithm to achieve adaptive policies via a simple heuristic composition of a set of fixed policies. Experiments on Chinese -> English and German -> English show that our adaptive policies can outperform fixed ones by up to 4 BLEU points for the same latency, and more surprisingly, it even surpasses the BLEU score of full-sentence translation in the greedy mode (and very close to beam mode), but with much lower latency.

Session 4B Semantics: Lexical-2

Breaking Through the 80% Glass Ceiling: Raising the State of the Art in Word Sense Disambiguation by Incorporating Knowledge Graph Information [Website][PDF]
Michele Bevilacqua and Roberto Navigli 18:00–19:00

Neural architectures are the current state of the art in Word Sense Disambiguation (WSD). However, they make limited use of the vast amount of relational information encoded in Lexical Knowledge Bases (LKB). We present Enhanced WSD Integrating Synset Embeddings and Relations (EWISER), a neural supervised architecture that is able to tap into this wealth of knowledge by embedding information from the LKB graph within the neural architecture, and to exploit pretrained synset embeddings, enabling the network to predict synsets that are not in the training set. As a result, we set a new state of the art on almost all the evaluation settings considered, also breaking through, for the first time, the 80% ceiling on the concatenation of all the standard all-words English WSD evaluation benchmarks. On multilingual all-words WSD, we report state-of-the-art results by training on nothing but English.

Glyph2Vec: Learning Chinese Out-of-Vocabulary Word Embedding from Glyphs [Website][PDF]
Hong-You Chen, SZ-HAN YU, and Shou-de Lin 18:00–19:00

Chinese NLP applications that rely on large text often contain huge amounts of vocabulary which are sparse in corpus. We show that characters' written form, *Glyphs*, in ideographic languages could carry rich semantics. We present a multi-modal model, *Glyph2Vec*, to tackle Chinese out-of-vocabulary word embedding problem. Glyph2Vec extracts visual features from word glyphs to expand current word embedding space for out-of-vocabulary word embedding, without the need of accessing any corpus, which is useful for improving Chinese NLP systems, especially for low-resource scenarios. Experiments across different applications show the significant effectiveness of our model.

[TACL] Learning Lexical Subspaces in a Distributional Vector Space [Website][PDF]
Kushal Arora, Aishik Chakraborty, and Jackie Chi Kit Cheung 18:00–19:00

In this paper, we propose LexSub, a novel approach towards unifying lexical and distributional semantics. We inject knowledge about lexical-semantic relations into distributional word embeddings by defining subspaces of the distributional vector space in which a lexical relation should hold. Our framework can handle symmetric attract and repel relations (e.g., synonymy and antonymy, respectively), as well as asymmetric relations (e.g., hypernymy and meronomy). In a suite of intrinsic benchmarks, we show that our model outperforms previous approaches on relatedness tasks and on hypernymy classification and detection, while being competitive on word similarity tasks. It also outperforms previous systems on extrinsic classification tasks that benefit from exploiting lexical relational cues. We perform a series of analyses to understand the behaviors of our model.

Moving Down the Long Tail of Word Sense Disambiguation with Gloss Informed Bi-encoders [Website][PDF]
Terra Blevins and Luke Zettlemoyer 18:00–19:00

A major obstacle in Word Sense Disambiguation (WSD) is that word senses are not uniformly distributed, causing existing models to generally perform poorly on senses that are either rare or unseen during training. We propose a bi-encoder model that independently embeds (1) the target word with its surrounding context and (2) the dictionary definition, or gloss, of each sense. The encoders are jointly optimized in the same representation space, so that sense disambiguation can be performed by finding the nearest sense embedding for each target word embedding. Our system outperforms previous state-of-the-art models on English all-words WSD; these gains predominantly come from improved performance on rare senses, leading to a 31.1% error reduction on less frequent senses over prior work. This demonstrates that rare senses can be more effectively disambiguated by modeling their definitions.

Multidirectional Associative Optimization of Function-Specific Word Representations [Website][PDF]
Daniela Gerz, Ivan Vulić, Marek Rei, Roi Reichart, and Anna Korhonen 18:00–19:00

We present a neural framework for learning associations between interrelated groups of words such as the ones found in Subject-Verb-Object (SVO) structures. Our model induces a joint function-specific word vector space, where vectors of e.g. plausible SVO compositions lie close together. The model retains information about word group membership even in the joint space, and can thereby effectively be applied to a number of tasks reasoning over the SVO structure. We show the robustness and versatility of the proposed framework by reporting state-of-the-art results on the tasks of estimating selectional preference and event similarity. The results indicate that the combinations of representations learned with our task-independent model outperform task-specific architectures from prior work, while reducing the number of parameters by up to 95%.

Predicting Degrees of Technicality in Automatic Terminology Extraction [Website][PDF]
Anna Hätty, Dominik Schlechtweg, Michael Dorna, and Sabine Schulte im Walde 18:00–19:00

While automatic term extraction is a well-researched area, computational approaches to distinguish between degrees of technicality are still understudied. We semi-automatically create a German gold standard of technicality across four domains, and illustrate the impact of a web-crawled general-language corpus on technicality prediction. When defining a classification approach that combines general-language and domain-specific word embeddings, we go beyond previous work and align vector spaces to gain comparative embeddings. We suggest two novel models to exploit general- vs. domain-specific comparisons: a simple neural network model with pre-computed comparative-embedding information as input, and a multi-channel model computing the comparison internally. Both models outperform previous approaches, with the multi-channel model performing best.

Verbal Multiword Expressions for Identification of Metaphor [Website][PDF]
Omid Rohanian, Marek Rei, Shiva Taslimipoor, and Le An Ha 18:00–19:00

Metaphor is a linguistic device in which a concept is expressed by mentioning another. Identifying metaphorical expressions, therefore, requires a non-compositional understanding of semantics. Multiword Expressions (MWEs), on the other hand, are linguistic phenomena with varying degrees of semantic opacity and their identification poses a challenge to computational models. This work is the first attempt at analysing the interplay of metaphor and MWEs processing through the design of a neural architecture whereby classification of metaphors is enhanced by informing the model of the presence of MWEs. To the best of our knowledge, this is the first "MWE-aware" metaphor identification system paving the way for further experiments on the complex interactions of these phenomena. The results and analyses show that this proposed architecture reach state-of-the-art on two different established metaphor datasets.

Session 4B: Student Research Workshop

A Geometry-Inspired Attack for Generating Natural Language Adversarial Examples [Website]
Zhao Meng and Roger Wattenhofer 18:00–19:00

Generating adversarial examples for natural language is hard, as natural language consists of discrete symbols and examples are often of variable lengths. In this paper, we propose a geometry-inspired attack for generating natural language adversarial examples. Our attack generates adversarial examples by iteratively approximating the decision boundary of deep neural networks. Experiments on two datasets with two different models show that our attack fools the models with high success rates, while only replacing a few words. Human evaluation shows that adversarial examples generated by our attack are hard for humans to recognize. Further experiments show that adversarial training can improve model robustness against our attack.

Effectively Aligning and Filtering Parallel Corpora under Sparse Data Conditions [Website][PDF]
Steinþór Steingrímsson, Hrafn Loftsson, and Andy Way 18:00–19:00

Parallel corpora are key to developing good machine translation systems. However, abundant parallel data are hard to come by, especially for languages with a low number of speakers. When rich morphology exacerbates the data sparsity problem, it is imperative to have accurate alignment and filtering methods that can help make the most of what is available by maximising the number of correctly translated segments in a corpus and minimising noise by removing incorrect translations and segments containing extraneous data. This paper sets out a research plan for improving alignment and filtering methods for parallel texts in low-resource settings. We propose an effective unsupervised alignment method to tackle the alignment problem. Moreover, we propose a strategy to supplement state-of-the-art models with automatically extracted information using basic NLP tools to effectively handle rich morphology.

Understanding Points of Correspondence between Sentences for Abstractive Summarization [Website][PDF]
Logan Lebanoff, John Muchovej, Franck Dernoncourt, Doo Soon Kim, Lidan Wang, Walter Chang, and Fei Liu 18:00–19:00

Fusing sentences containing disparate content is a remarkable human ability that helps create informative and succinct summaries. Such a simple task for humans has remained challenging for modern abstractive summarizers, substantially restricting their applicability in real-world scenarios. In this paper, we present an investigation into fusing sentences drawn from a document by introducing the notion of points of correspondence, which are cohesive devices that tie any two sentences together into a coherent text. The types of points of correspondence are delineated by text cohesion theory, covering pronominal and nominal referencing, repetition and beyond. We create a dataset containing the documents, source and fusion sentences, and human annotations of points of correspondence between sentences. Our dataset bridges the gap between coreference resolution and summarization. It is publicly shared to serve as a basis for future work to measure the success of sentence fusion systems.

Noise-Based Augmentation Techniques for Emotion Datasets: What do we Recommend? [Website]
Mimansa Jaiswal and Emily Mower Provost 18:00–19:00

Emotion recognition systems are widely used for many downstream applications such as mental health monitoring, educational problems diagnosis, hate speech classification and targeted advertising. Yet, these systems are generally trained on audio or multimodal datasets collected in a laboratory environment.While acoustically different, they are generally free of major environmental noises. The result is that systems trained on these datasets falter when presented with noisy data, even when that noise doesn't affect the human perception of emotions. In this work, we use multiple categories of environmental and synthetic noises to generate black box adversarial examples and use these noises to modify the samples in the IEMOCAP dataset. We evaluate how both human and machine emotion perception changes when noise is introduced. We find that the trained state-of-the-art models fail to classify even moderately noisy samples that humans usually have no trouble comprehend-ing, demonstrating the brittleness of these systems in real world conditions.

Session 4B: Summarization-3

Attend to Medical Ontologies: Content Selection for Clinical Abstractive Summarization [Website][PDF]
Sajad Sotudeh Gharebagh, Nazli Goharian, and Ross Filice 18:00–19:00

Sequence-to-sequence (seq2seq) network is a well-established model for text summarization task. It can learn to produce readable content; however, it falls short in effectively identifying key regions of the source. In this paper, we approach the content selection problem for clinical abstractive summarization by augmenting salient ontological terms into the summarizer. Our experiments on two publicly available clinical data sets (107,372 reports of MIMIC-CXR, and 3,366 reports of OpenI) show that our model statistically significantly boosts state-of-the-art results in terms of ROUGE metrics (with improvements: 2.9% RG-1, 2.5% RG-2, 1.9% RG-L), in the healthcare domain where any range of improvement impacts patients' welfare.

On Faithfulness and Factuality in Abstractive Summarization [Website][PDF]
Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald 18:00–19:00

It is well known that the standard likelihood training and approximate decoding objectives in neural text generation models lead to less human-like responses for open-ended tasks such as language modeling and story generation. In this paper we have analyzed limitations of these models for abstractive document summarization and found that these models are highly prone to hallucinate content that is unfaithful to the input document. We conducted a large scale human evaluation of several neural abstractive summarization systems to better understand the types of hallucinations they produce. Our human annotators found substantial amounts of hallucinated content in all model generated summaries. However, our analysis does show that pretrained models are better summarizers not only in terms of raw metrics, i.e., ROUGE, but also in generating faithful and factual summaries as evaluated by humans. Furthermore, we show that textual entailment measures better correlate with faithfulness than standard metrics, potentially leading the way to automatic evaluation metrics as well as training and decoding criteria.

Screenplay Summarization Using Latent Narrative Structure [Website][PDF]
Pinelopi Papalampidi, Frank Keller, Lea Frermann, and Mirella Lapata 18:00–19:00

Most general-purpose extractive summarization models are trained on news articles, which are short and present all important information upfront. As a result, such models are biased on position and often perform a smart selection of sentences from the beginning of the document. When summarizing long narratives, which have complex structure and present information piecemeal, simple position heuristics are not sufficient. In this paper, we propose to explicitly incorporate the underlying structure of narratives into general unsupervised and supervised extractive summarization models. We formalize narrative structure in terms of key narrative events (turning points) and treat it as latent in order to summarize screenplays (i.e., extract an optimal sequence of scenes). Experimental results on the CSI corpus of TV screenplays, which we augment with scene-level summarization labels, show that latent turning points correlate with important aspects of a CSI episode and improve summarization performance over general extractive algorithms leading to more complete and diverse summaries.

Unsupervised Opinion Summarization with Noising and Denoising [Website][PDF]
Reinald Kim Amplayo and Mirella Lapata 18:00–19:00

The supervised training of high-capacity models on large datasets containing hundreds of thousands of document-summary pairs is critical to the recent success of deep learning techniques for abstractive summarization. Unfortunately, in most domains (other than news) such training data is not available and cannot be easily sourced. In this paper we enable the use of supervised learning for the setting where there are only documents available (e.g., product or business reviews) without ground truth summaries. We create a synthetic dataset from a corpus of user reviews by sampling a review, pretending it is a summary, and generating noisy versions thereof which we treat as pseudo-review input. We introduce several linguistically motivated noise generation functions and a summarization model which learns to denoise the input and generate the original review. At test time, the model accepts genuine reviews and generates a summary containing salient opinions, treating those that do not reach consensus as noise. Extensive automatic and human evaluation shows that our model brings substantial improvements over both abstractive and extractive baselines.

Demo Session 4C

Time: *18:30–19:15*

Xiaomingbot: A Multilingual Robot News Reporter [Website][PDF]
Runxin Xu, Jun Cao, Mingxuan Wang, Jiaze Chen, Hao Zhou, Ying Zeng, Yuping Wang, Li Chen, Xiang Yin, Xijin Zhang, Songcheng Jiang, Yuxuan Wang, and Lei Li

This paper proposes the building of Xiaomingbot, an intelligent, multilingual and multimodal software robot equipped with four inte- gral capabilities: news generation, news translation, news reading and avatar animation. Its system summarizes Chinese news that it automatically generates from data tables. Next, it translates the summary or the full article into multiple languages, and reads the multi- lingual rendition through synthesized speech. Notably, Xiaomingbot utilizes a voice cloning technology to synthesize the speech trained from a real person's voice data in one input language. The proposed system enjoys several merits: it has an animated avatar, and is able to generate and read multilingual news. Since it was put into practice, Xiaomingbot has written over 600,000 articles, and gained over 150,000 followers on social media platforms.

jiant: A Software Toolkit for Research on General-Purpose Text Understanding Models [Website][PDF]
Yada Pruksachatkun, Phil Yeres, Haokun Liu, Jason Phang, Phu Mon Htut, Alex Wang, Ian Tenney, and Samuel R. Bowman

We introduce jiant, an open source toolkit for conducting multitask and transfer learning experiments on English NLU tasks. jiant enables modular and configuration driven experimentation with state-of-the-art models and a broad set of tasks for probing, transfer learning, and multitask training experiments. jiant implements over 50 NLU tasks, including all GLUE and SuperGLUE benchmark tasks. We demonstrate that jiant reproduces published performance on a variety of tasks and models, e.g., RoBERTa and BERT.

The Microsoft Toolkit of Multi-Task Deep Neural Networks for Natural Language Understanding [Website][PDF]
Xiaodong Liu, Yu Wang, Jianshu Ji, Hao Cheng, Xueyun Zhu, Emmanuel Awa, Pengcheng He, Weizhu Chen, Hoifung Poon, Guihong Cao, and Jianfeng Gao

We present MT-DNN, an open-source natural language understanding (NLU) toolkit that makes it easy for researchers and developers to train customized deep learning models. Built upon PyTorch and Transformers, MT-DNN is designed to facilitate rapid customization for a broad spectrum of NLU tasks, using a variety of objectives (classification, regression, structured prediction) and text encoders (e.g., RNNs, BERT, RoBERTa, UniLM). A unique feature of MT-DNN is its built-in support for robust and transferable learning using the adversarial multi-task learning paradigm. To enable efficient production deployment, MT-DNN supports multi-task knowledge distillation, which can substantially compress a deep neural model without significant performance drop. We demonstrate the effectiveness of MT-DNN on a wide range of NLU applications across general and biomedical domains. The software and pre-trained models will be publicly available at https://github.com/namisan/mt-dnn.

Demo Session 5A

Time: *20:00–20:45*

GAIA: A Fine-grained Multimedia Knowledge Extraction System [Website][PDF]
Manling Li, Alireza Zareian, Ying Lin, Xiaoman Pan, Spencer Whitehead, Brian Chen, Bo Wu, Heng Ji, Shih-Fu Chang, Clare Voss, Daniel Napierski, and Marjorie Freedman

We present the first comprehensive, open source multimedia knowledge extraction system that takes a massive stream of unstructured, heterogeneous multimedia data from various sources and languages as input, and creates a coherent, structured knowledge base, indexing entities, relations, and events, following a rich, fine-grained ontology. Our system, GAIA, enables seamless search of complex graph queries, and retrieves multimedia evidence including text, images and videos. GAIA achieves top performance at the recent NIST TAC SM-KBP2019 evaluation. The system is publicly available at GitHub and DockerHub, with a narrated video that documents the system.

Trialstreamer: Mapping and Browsing Medical Evidence in Real-Time [Website][PDF]
Benjamin Nye, Ani Nenkova, Iain Marshall, and Byron C. Wallace

We introduce Trialstreamer, a living database of clinical trial reports. Here we mainly describe the evidence extraction component; this extracts from biomedical abstracts key pieces of information that clinicians need when appraising the literature, and also the relations between these. Specifically, the system extracts descriptions of trial participants, the treatments compared in each arm (the interventions), and which outcomes were measured. The system then attempts to infer which interventions were reported to work best by determining their relationship with identified trial outcome measures. In addition to summarizing individual trials, these extracted data elements allow automatic synthesis of results across many trials on the same topic. We apply the system at scale to all reports of randomized controlled trials indexed in MEDLINE, powering the automatic generation of evidence maps, which provide a global view of the efficacy of different interventions combining data from all relevant clinical trials on a topic. We make all code and models freely available alongside a demonstration of the web interface.

pyBART: Evidence-based Syntactic Transformations for IE [Website][PDF]
Aryeh Tiktinsky, Yoav Goldberg, and Reut Tsarfaty

Syntactic dependencies can be predicted with high accuracy, and are useful for both machine-learned and pattern-based information extraction tasks. However, their utility can be improved. These syntactic dependencies are designed to accurately reflect syntactic relations, and they do not make semantic relations explicit. Therefore, these representations lack many explicit connections between content words, that would be useful for downstream applications. Proposals like English Enhanced UD improve the situation by extending universal dependency trees with additional explicit arcs. However, they are not available to Python users, and are also limited in coverage. We introduce a broad-coverage, data-driven and linguistically sound set of transformations, that makes event-structure and many lexical relations explicit. We present pyBART, an easy-to-use open-source Python library for converting English UD trees either to Enhanced UD graphs or to our representation. The library can work as a standalone package or be integrated within a spaCy NLP pipeline. When evaluated in a pattern-based relation extraction scenario, our representation results in higher extraction scores than Enhanced UD, while requiring fewer patterns.

Session 5A Overview – Monday, July 6, 2020 20:00–21:00

Track A *Dialogue and Interactive Systems-9* Abstracts	Grounded Conversation Generation as Guided Traverses in Commonsense Knowledge Graphs *Zhang, Liu, Xiong, and Liu* [Website][PDF]	Learning an Unreferenced Metric for Online Dialogue Evaluation *Sinha, Parthasarathi, Wang, Lowe, Hamilton, and Pineau* [Website][PDF]	Negative Training for Neural Dialogue Response Generation *He and Glass* [Website][PDF]	Recursive Template-based Frame Generation for Task Oriented Dialog *Gangadharaiah and Narayanaswamy* [Website][PDF]	Speak to your Parser: Interactive Text-to-SQL with Natural Language Feedback *Elgohary, Hosseini, and Hassan Awadallah* [Website][PDF]
Track B *Generation-8* Abstracts	Automatic Detection of Generated Text is Easiest when Humans are Fooled *Ippolito, Duckworth, Callison-Burch, and Eck* [Website][PDF]	Automatic Poetry Generation from Prosaic Text *Van de Cruys* [Website][PDF]	Bridging the Structural Gap Between Encoding and Decoding for Data-To-Text Generation *Zhao, Walker, and Chaturvedi* [Website][PDF]	Cross-modal Language Generation using Pivot Stabilization for Web-scale Language Coverage *Thapliyal and Soricut* [Website][PDF]	Enabling Language Models to Fill in the Blanks *Donahue, Lee, and Liang* [Website][PDF]
	Few-Shot NLG with Pre-Trained Language Model *Chen, Eavani, Chen, Liu, and Wang* [Website][PDF]	INSET: Sentence Infilling with INter-SEntential Transformer *Huang, Zhang, Elachqar, and Cheng* [Website][PDF]	Improved Natural Language Generation via Loss Truncation *Kang and Hashimoto* [Website][PDF]	Improving Adversarial Text Generation by Modeling the Distant Future *Zhang, Chen, Gan, Wang, Shen, Wang, Wen, and Carin* [Website][PDF]	Learning to Update Natural Language Comments Based on Code Changes *Panthaplackel, Nie, Gligoric, Li, and Mooney* [Website][PDF]
	Politeness Transfer: A Tag and Generate Approach *Madaan, Setlur, Parekh, Poczos, Neubig, Yang, Salakhutdinov, Black, and Prabhumoye* [Website][PDF]	Reverse Engineering Configurations of Neural Text Generation Models *Tay, Bahri, Zheng, Brunk, Metzler, and Tomkins* [Website][PDF]	Simple and Effective Retrieve-Edit-Rerank Text Generation *Hossain, Ghazvininejad, and Zettlemoyer* [Website][PDF]		
Track C *Information Retrieval and Text Mining-5* Abstracts	Contextualized Weak Supervision for Text Classification *Mekala and Shang* [Website][PDF]				
Track D *Machine Learning for NLP-3* Abstracts	Calibrating Structured Output Predictors for Natural Language Processing *Jagannatha and* [Website][PDF]	Active Imitation Learning with Noisy Guidance *Brantley, Daumé III, and Sharaf* [Website][PDF]	ExpBERT: Representation Engineering with Natural Language Explanations *Murty, Koh, and Liang* [Website][PDF]	GAN-BERT: Generative Adversarial Learning for Robust Text Classification with a Bunch of Labeled Examples *Croce, Castellucci, and Basili* [Website][PDF]	Generalizing Natural Language Analysis through Span-relation Representations *Jiang, Xu, Araki, and Neubig* [Website][PDF]

	Learning to Contextually Aggregate Multi-Source Supervision for Sequence Labeling *Lan, Huang, Lin, Jiang, Liu, and Ren* [Website][PDF]	MixText: Linguistically-Informed Interpolation of Hidden Space for Semi-Supervised Text Classification *Chen, Yang, and Yang* [Website][PDF]	MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices *Sun, Yu, Song, Liu, Yang, and Zhou* [Website][PDF]	On Importance Sampling-Based Evaluation of Latent Language Models *Logan IV, Gardner, and Singh* [Website][PDF]	SMART: Robust and Efficient Fine-Tuning for Pre-trained Natural Language Models through Principled Regularized Optimization *Jiang, He, Chen, Liu, Gao, and Zhao* [Website][PDF]
	Stolen Probability: A Structural Weakness of Neural Language Models *Demeter, Kimmel, and Downey* [Website][PDF]	Taxonomy Construction of Unseen Domains via Graph-based Cross-Domain Knowledge Transfer *Shang, Dash, Chowdhury, Mihindukulasooriya, and Gliozzo* [Website][PDF]	To Pretrain or Not to Pretrain: Examining the Benefits of Pretrainng on Resource Rich Tasks *Wang, Khabsa, and Ma* [Website][PDF]	Why Overfitting Isn't Always Bad: Retrofitting Cross-Lingual Word Embeddings to Dictionaries *Zhang, Fujinuma, Paul, and Boyd-Graber* [Website][PDF]	XtremeDistil: Multi-stage Distillation for Massive Multilingual Models *Mukherjee and Hassan Awadallah* [Website][PDF]
Track E *Machine Translation-6* Abstracts	ENGINE: Energy-Based Inference Networks for Non-Autoregressive Machine Translation *Tu, Pang, Wiseman, and Gimpel* [Website][PDF]	Improving Non-autoregressive Neural Machine Translation with Monolingual Data *Zhou and Keung* [Website][PDF]	Leveraging Monolingual Data with Self-Supervision for Multilingual Neural Machine Translation *Siddhant, Bapna, Cao, Firat, Chen, Kudugunta, Arivazhagan, and Wu* [Website][PDF]	Location Attention for Extrapolation to Longer Sequences *Dubois, Dagan, Hupkes, and Bruni* [Website][PDF]	On The Evaluation of Machine Translation SystemsTrained With Back-Translation *Edunov, Ott, Ranzato, and Auli* [Website][PDF]
	Opportunistic Decoding with Timely Correction for Simultaneous Translation *Zheng, Ma, Zheng, Liu, and Huang* [Website][PDF]	Simultaneous Translation Policies: From Fixed to Adaptive *Zheng, Liu, Zheng, Ma, Liu, and Huang* [Website][PDF]			
Track F *Textual Inference and Other Areas of Semantics-2* Abstracts	Can We Predict New Facts with Open Knowledge Graph Embeddings? A Benchmark for Open Link Prediction *Broscheit, Gashteovski, Wang, and Gemulla* [Website][PDF]	[TACL] Decomposing Generalization: Models of Generic, Habitual and Episodic Statements *Govindarajan, Durme, and White* [Website][PDF]	INFOTABS: Inference on Tables as Semi-structured Data *Gupta, Mehta, Nokhiz, and Srikumar* [Website][PDF]	[TACL] Inherent Disagreements in Human Textual Inferences *Pavlick and Kwiatkowski* [Website][PDF]	Interactive Machine Comprehension with Information Seeking Agents *Yuan, Fu, Côté, Tay, Pal, and Trischler* [Website][PDF]
	Syntactic Data Augmentation Increases Robustness to Inference Heuristics *Min, McCoy, Das, Pitler, and Linzen* [Website][PDF]				

Track G *Speech and Multimodality-2* Abstracts	[TACL] Acoustic-Prosodic and Lexical Cues to Deception and Trust: Deciphering How People Detect Lies *Chen, Levitan, Levine, Mandic, and Hirschberg* [Website] [PDF]	Improved Speech Representations with Multi-Target Autoregressive Predictive Coding *Chung and Glass* [Website] [PDF]	Integrating Multimodal Information in Large Pretrained Transformers *Rahman, Hasan, Lee, Bagher Zadeh, Mao, Morency, and Hoque* [Website] [PDF]	MultiQT: Multimodal learning for real-time question tracking in speech *D. Havtorn, Latko, Edin, Maaløe, Borgholt, Belgrano, Jacobsen, Sdun, and Agić* [Website] [PDF]	Multimodal and Multiresolution Speech Recognition with Transformers *Paraskevopoulos, Parthasarathy, Khare, and Sundaram* [Website] [PDF]
	Phone Features Improve Speech Translation *Salesky and Black* [Website] [PDF]				
Track H *Theory and Formalism in NLP (Linguistic and Mathematical)-3* Abstracts	A Formal Hierarchy of RNN Architectures *Merrill, Weiss, Goldberg, Schwartz, Smith, and Yahav* [Website] [PDF]	Emergence of Syntax Needs Minimal Supervision *Bailly and Gábor* [Website] [PDF]			

Session 5A Details

Session 5A: Dialogue and Interactive Systems-9

Grounded Conversation Generation as Guided Traverses in Commonsense Knowledge Graphs [Website][PDF]
Houyu Zhang, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu 20:00–21:00

Human conversations naturally evolve around related concepts and hop to distant concepts. This paper presents a new conversation generation model, ConceptFlow, which leverages commonsense knowledge graphs to explicitly model conversation flows. By grounding conversations to the concept space, ConceptFlow represents the potential conversation flow as traverses in the concept space along commonsense relations. The traverse is guided by graph attentions in the concept graph, moving towards more meaningful directions in the concept space, in order to generate more semantic and informative responses. Experiments on Reddit conversations demonstrate ConceptFlow's effectiveness over previous knowledge-aware conversation models and GPT-2 based models while using 70% fewer parameters, confirming the advantage of explicit modeling conversation structures. All source codes of this work are available at https://github.com/thunlp/ConceptFlow.

Learning an Unreferenced Metric for Online Dialogue Evaluation [Website][PDF]
Koustuv Sinha, Prasanna Parthasarathi, Jasmine Wang, Ryan Lowe, William L. Hamilton, and Joelle Pineau 20:00–21:00

Evaluating the quality of a dialogue interaction between two agents is a difficult task, especially in open-domain chit-chat style dialogue. There have been recent efforts to develop automatic dialogue evaluation metrics, but most of them do not generalize to unseen datasets and/or need a human-generated reference response during inference, making it infeasible for online evaluation. Here, we propose an unreferenced automated evaluation metric that uses large pre-trained language models to extract latent representations of utterances, and leverages the temporal transitions that exist between them. We show that our model achieves higher correlation with human annotations in an online setting, while not requiring true responses for comparison during inference.

Negative Training for Neural Dialogue Response Generation [Website][PDF]
Tianxing He and James Glass 20:00–21:00

Although deep learning models have brought tremendous advancements to the field of open-domain dialogue response generation, recent research results have revealed that the trained models have undesirable generation behaviors, such as malicious responses and generic (boring) responses. In this work, we propose a framework named "Negative Training" to minimize such behaviors. Given a trained model, the framework will first find generated samples that exhibit the undesirable behavior, and then use them to feed negative training signals for fine-tuning the model. Our experiments show that negative training can significantly reduce the hit rate of malicious responses, or discourage frequent responses and improve response diversity.

Recursive Template-based Frame Generation for Task Oriented Dialog [Website][PDF]
Rashmi Gangadharaiah and Balakrishnan Narayanaswamy 20:00–21:00

The Natural Language Understanding (NLU) component in task oriented dialog systems processes a user's request and converts it into structured information that can be consumed by downstream components such as the Dialog State Tracker (DST). This information is typically represented as a semantic frame that captures the intent and slot-labels provided by the user. We first show that such a shallow representation is insufficient for complex dialog scenarios, because it does not capture the recursive nature inherent in many domains. We propose a recursive, hierarchical frame-based representation and show how to learn it from data. We formulate the frame generation task as a template-based tree decoding task, where the decoder recursively generates a template and then fills slot values into the template. We extend local tree-based loss functions with terms that provide global supervision and show how to optimize them end-to-end. We achieve a small improvement on the widely used ATIS dataset and a much larger improvement on a more complex dataset we describe here.

Speak to your Parser: Interactive Text-to-SQL with Natural Language Feedback [Website][PDF]
Ahmed Elgohary, saghar Hosseini, and Ahmed Hassan Awadallah 20:00–21:00

We study the task of semantic parse correction with natural language feedback. Given a natural language utterance, most semantic parsing systems pose the problem as one-shot translation where the utterance is mapped to a corresponding logical form. In this paper, we investigate a more interactive scenario where humans can further interact with the system by providing free-form natural language feedback to correct the system when it generates an inaccurate interpretation of an initial utterance. We focus on natural language to SQL systems and construct, SPLASH, a dataset of utterances, incorrect SQL interpretations and the corresponding natural language feedback. We compare various reference models for the correction task and show that incorporating such a rich form of feedback can significantly improve the overall semantic parsing accuracy while retaining the flexibility of natural language interaction. While we estimated human correction accuracy is 81.5%, our best model achieves only 25.1%, which leaves a large gap for improvement in future research. SPLASH is publicly available at https://aka.ms/Splash_dataset.

Session 5A: Generation-8

Automatic Detection of Generated Text is Easiest when Humans are Fooled [Website][PDF]
Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck 20:00–21:00

Recent advancements in neural language modelling make it possible to rapidly generate vast amounts of human-sounding text. The capabilities of humans and automatic discriminators to detect machine-generated text have been a large source of research interest, but humans and machines rely on different cues to make their decisions. Here, we perform careful benchmarking and analysis of three popular sampling-based decoding strategies—top-_k_, nucleus sampling, and untruncated random sampling—and show that improvements in decoding methods have primarily optimized for fooling humans. This comes at the expense of introducing statistical abnormalities that make detection easy for automatic systems. We also show that though both human and automatic detector performance improve with longer excerpt length, even multi-sentence excerpts can fool expert human raters over 30% of the time. Our findings reveal the importance of using both human and automatic detectors to assess the humanness of text generation systems.

Automatic Poetry Generation from Prosaic Text [Website][PDF]
Tim Van de Cruys 20:00–21:00

In the last few years, a number of successful approaches have emerged that are able to adequately model various aspects of natural language. In particular, language models based on neural networks have improved the state of the art with regard to predictive language modeling, while topic models are successful at capturing clear-cut, semantic dimensions. In this paper, we will explore how these approaches can be adapted and combined to model the linguistic and literary aspects needed for poetry generation. The system is exclusively trained on standard, non-poetic text, and its output is constrained in order to confer a poetic character to the generated verse. The framework is applied to the generation of poems in both English and French, and is equally evaluated for both languages. Even though it only uses standard, non-poetic text as input, the system yields state of the art results for poetry generation.

Bridging the Structural Gap Between Encoding and Decoding for Data-To-Text Generation [Website][PDF]
Chao Zhao, Marilyn Walker, and Snigdha Chaturvedi 20:00–21:00

Generating sequential natural language descriptions from graph-structured data (e.g., knowledge graph) is challenging, partly because of the structural differences between the input graph and the output text. Hence, popular sequence-to-sequence models, which require serialized input, are not a natural fit for this task. Graph neural networks, on the other hand, can better encode the input graph but broaden the structural gap between the encoder and decoder, making faithful generation difficult. To narrow this gap, we propose DualEnc, a dual encoding model that can not only incorporate the graph structure, but can also cater to the linear structure of the output text. Empirical comparisons with strong single-encoder baselines demonstrate that dual encoding can significantly improve the quality of the generated text.

Cross-modal Language Generation using Pivot Stabilization for Web-scale Language Coverage [Website][PDF]
Ashish V. Thapliyal and Radu Soricut 20:00–21:00

Cross-modal language generation tasks such as image captioning are directly hurt in their ability to support non-English languages by the trend of data-hungry models combined with the lack of non-English annotations. We investigate potential solutions for combining existing language-generation annotations in English with translation capabilities in order to create solutions at web-scale in both domain and language coverage. We describe an approach called Pivot-Language Generation Stabilization (PLuGS), which leverages directly at training time both existing English annotations (gold data) as well as their machine-translated versions (silver data); at run-time, it generates first an English caption and then a corresponding target-language caption. We show that PLuGS models outperform other candidate solutions in evaluations performed over 5 different target languages, under a large-domain testset using images from the Open Images dataset. Furthermore, we find an interesting effect where the English captions generated by the PLuGS models are better than the captions generated by the original, monolingual English model.

Enabling Language Models to Fill in the Blanks [Website][PDF]
Chris Donahue, Mina Lee, and Percy Liang 20:00–21:00

We present a simple approach for *text infilling*, the task of predicting missing spans of text at any position in a document. While infilling could enable rich functionality especially for writing assistance tools, more attention has been devoted to language modeling—a special case of infilling where text is predicted at the end of a document. In this paper, we aim to extend the capabilities of language models (LMs) to the more general task of infilling. To this end, we train (or fine tune) off-the-shelf LMs on sequences containing the concatenation of artificially-masked text and the text which was masked. We show that this approach, which we call *infilling by language modeling*, can enable LMs to infill entire sentences effectively on three different domains: short stories, scientific abstracts, and lyrics. Furthermore, we show that humans have difficulty identifying sentences infilled by our approach as machine-generated in the domain of short stories.

Few-Shot NLG with Pre-Trained Language Model [Website][PDF]
Zhiyu Chen, Harini Eavani, Wenhu Chen, Yinyin Liu, and William Yang Wang 20:00–21:00

Neural-based end-to-end approaches to natural language generation (NLG) from structured data or knowledge are data-hungry, making their adoption for real-world applications difficult with limited data. In this work, we propose

the new task of few-shot natural language generation. Motivated by how humans tend to summarize tabular data, we propose a simple yet effective approach and show that it not only demonstrates strong performance but also provides good generalization across domains. The design of the model architecture is based on two aspects: content selection from input data and language modeling to compose coherent sentences, which can be acquired from prior knowledge. With just 200 training examples, across multiple domains, we show that our approach achieves very reasonable performances and outperforms the strongest baseline by an average of over 8.0 BLEU points improvement. Our code and data can be found at https://github.com/czyssrs/Few-Shot-NLG

INSET: Sentence Infilling with INter-SEntential Transformer [Website][PDF]
Yichen Huang, Yizhe Zhang, Oussama Elachqar, and Yu Cheng 20:00–21:00

Missing sentence generation (or sentence in-filling) fosters a wide range of applications in natural language generation, such as document auto-completion and meeting note expansion. This task asks the model to generate intermediate missing sentences that can syntactically and semantically bridge the surrounding context. Solving the sentence infilling task requires techniques in natural language processing ranging from understanding to discourse-level planning to generation. In this paper, we propose a framework to decouple the challenge and address these three aspects respectively, leveraging the power of existing large-scale pre-trained models such as BERT and GPT-2. We empirically demonstrate the effectiveness of our model in learning a sentence representation for generation and further generating a missing sentence that fits the context.

Improved Natural Language Generation via Loss Truncation [Website][PDF]
Daniel Kang and Tatsunori Hashimoto 20:00–21:00

Neural language models are usually trained to match the distributional properties of large-scale corpora by minimizing the log loss. While straightforward to optimize, this approach forces the model to reproduce all variations in the dataset, including noisy and invalid references (e.g., misannotations and hallucinated facts). Even a small fraction of noisy data can degrade the performance of log loss. As an alternative, prior work has shown that minimizing the distinguishability of generated samples is a principled and robust loss that can handle invalid references. However, distinguishability has not been used in practice due to challenges in optimization and estimation. We propose loss truncation: a simple and scalable procedure which adaptively removes high log loss examples as a way to optimize for distinguishability. Empirically, we demonstrate that loss truncation outperforms existing baselines on distinguishability on a summarization task. Furthermore, we show that samples generated by the loss truncation model have factual accuracy ratings that exceed those of baselines and match human references.

Improving Adversarial Text Generation by Modeling the Distant Future [Website][PDF]
Ruiyi Zhang, Changyou Chen, Zhe Gan, Wenlin Wang, Dinghan Shen, Guoyin Wang, Zheng Wen, and Lawrence Carin 20:00–21:00

Auto-regressive text generation models usually focus on local fluency, and may cause inconsistent semantic meaning in long text generation. Further, automatically generating words with similar semantics is challenging, and hand-crafted linguistic rules are difficult to apply. We consider a text planning scheme and present a model-based imitation-learning approach to alleviate the aforementioned issues. Specifically, we propose a novel guider network to focus on the generative process over a longer horizon, which can assist next-word prediction and provide intermediate rewards for generator optimization. Extensive experiments demonstrate that the proposed method leads to improved performance.

Learning to Update Natural Language Comments Based on Code Changes [Website][PDF]
Sheena Panthaplackel, Pengyu Nie, Milos Gligoric, Junyi Jessy Li, and Raymond Mooney 20:00–21:00

We formulate the novel task of automatically updating an existing natural language comment based on changes in the body of code it accompanies. We propose an approach that learns to correlate changes across two distinct language representations, to generate a sequence of edits that are applied to the existing comment to reflect the source code modifications. We train and evaluate our model using a dataset that we collected from commit histories of open-source software projects, with each example consisting of a concurrent update to a method and its corresponding comment. We compare our approach against multiple baselines using both automatic metrics and human evaluation. Results reflect the challenge of this task and that our model outperforms baselines with respect to making edits.

Politeness Transfer: A Tag and Generate Approach [Website][PDF]
Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabas Poczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W Black, and Shrimai Prabhumoye 20:00–21:00

This paper introduces a new task of politeness transfer which involves converting non-polite sentences to polite sentences while preserving the meaning. We also provide a dataset of more than 1.39 instances automatically labeled for politeness to encourage benchmark evaluations on this new task. We design a tag and generate pipeline that identifies stylistic attributes and subsequently generates a sentence in the target style while preserving most of the source content. For politeness as well as five other transfer tasks, our model outperforms the state-of-the-art methods on automatic metrics for content preservation, with a comparable or better performance on style transfer accuracy. Additionally, our model surpasses existing methods on human evaluations for grammaticality, meaning preservation and transfer accuracy across all the six style transfer tasks. The data and code is located at https://github.com/tag-and-generate.

Reverse Engineering Configurations of Neural Text Generation Models [Website][PDF]
Yi Tay, Dara Bahri, Che Zheng, Clifford Brunk, Donald Metzler, and Andrew Tomkins 20:00–21:00

Recent advances in neural text generation modeling have resulted in a number of societal concerns related to how such approaches might be used in malicious ways. It is therefore desirable to develop a deeper understanding of the

fundamental properties of such models. The study of artifacts that emerge in machine generated text as a result of modeling choices is a nascent research area. To this end, the extent and degree to which these artifacts surface in generated text is still unclear. In the spirit of better understanding generative text models and their artifacts, we propose the new task of distinguishing which of several variants of a given model generated some piece of text. Specifically, we conduct an extensive suite of diagnostic tests to observe whether modeling choices (e.g., sampling methods, top-k probabilities, model architectures, etc.) leave detectable artifacts in the text they generate. Our key finding, which is backed by a rigorous set of experiments, is that such artifacts are present and that different modeling choices can be inferred by looking at generated text alone. This suggests that neural text generators may actually be more sensitive to various modeling choices than previously thought.

Simple and Effective Retrieve-Edit-Rerank Text Generation [Website] [PDF]
Nabil Hossain, Marjan Ghazvininejad, and Luke Zettlemoyer 20:00–21:00

Retrieve-and-edit seq2seq methods typically retrieve an output from the training set and learn a model to edit it to produce the final output. We propose to extend this framework with a simple and effective post-generation ranking approach. Our framework (i) retrieves several potentially relevant outputs for each input, (ii) edits each candidate independently, and (iii) re-ranks the edited candidates to select the final output. We use a standard editing model with simple task-specific re-ranking approaches, and we show empirically that this approach outperforms existing, significantly more complex methodologies. Experiments on two machine translation (MT) datasets show new state-of-art results. We also achieve near state-of-art performance on the Gigaword summarization dataset, where our analyses show that there is significant room for performance improvement with better candidate output selection in future work.

Session 5A: Information Retrieval and Text Mining-5

Contextualized Weak Supervision for Text Classification [Website][PDF]
Dheeraj Mekala and Jingbo Shang 20:00–21:00

Weakly supervised text classification based on a few user-provided seed words has recently attracted much attention from researchers. Existing methods mainly generate pseudo-labels in a context-free manner (e.g., string matching), therefore, the ambiguous, context-dependent nature of human language has been long overlooked. In this paper, we propose a novel framework ConWea, providing contextualized weak supervision for text classification. Specifically, we leverage contextualized representations of word occurrences and seed word information to automatically differentiate multiple interpretations of the same word, and thus create a contextualized corpus. This contextualized corpus is further utilized to train the classifier and expand seed words in an iterative manner. This process not only adds new contextualized, highly label-indicative keywords but also disambiguates initial seed words, making our weak supervision fully contextualized. Extensive experiments and case studies on real-world datasets demonstrate the necessity and significant advantages of using contextualized weak supervision, especially when the class labels are fine-grained.

Session 5A: Machine Learning for NLP-3

Calibrating Structured Output Predictors for Natural Language Processing [Website][PDF]
Abhyuday Jagannatha and hong yu hong 20:00–21:00

We address the problem of calibrating prediction confidence for output entities of interest in natural language processing (NLP) applications. It is important that NLP applications such as named entity recognition and question answering produce calibrated confidence scores for their predictions, especially if the applications are to be deployed in a safety-critical domain such as healthcare. However the output space of such structured prediction models are often too large to directly adapt binary or multi-class calibration methods. In this study, we propose a general calibration scheme for output entities of interest in neural network based structured prediction models. Our proposed method can be used with any binary class calibration scheme and a neural network model. Additionally, we show that our calibration method can also be used as an uncertainty-aware, entity-specific decoding step to improve the performance of the underlying model at no additional training cost or data requirements. We show that our method outperforms current calibration techniques for Named Entity Recognition, Part-of-speech tagging and Question Answering systems. We also observe an improvement in model performance from our decoding step across several tasks and benchmark datasets. Our method improves the calibration and model performance on out-of-domain test scenarios as well.

Active Imitation Learning with Noisy Guidance [Website][PDF]
Kianté Brantley, Hal Daumé III, and Amr Sharaf 20:00–21:00

Imitation learning algorithms provide state-of-the-art results on many structured prediction tasks by learning near-optimal search policies. Such algorithms assume training-time access to an expert that can provide the optimal action at any queried state; unfortunately, the number of such queries is often prohibitive, frequently rendering these approaches impractical. To combat this query complexity, we consider an active learning setting in which the learning algorithm has additional access to a much cheaper noisy heuristic that provides noisy guidance. Our algorithm, LEAQI, learns a difference classifier that predicts when the expert is likely to disagree with the heuristic, and queries the expert only when necessary. We apply LEAQI to three sequence labelling tasks, demonstrating significantly fewer queries to the expert and comparable (or better) accuracies over a passive approach.

ExpBERT: Representation Engineering with Natural Language Explanations [Website][PDF]
Shikhar Murty, Pang Wei Koh, and Percy Liang 20:00–21:00

Suppose we want to specify the inductive bias that married couples typically go on honeymoons for the task of extracting pairs of spouses from text. In this paper, we allow model developers to specify these types of inductive biases as natural language explanations. We use BERT fine-tuned on MultiNLI to "interpret" these explanations with respect to the input sentence, producing explanation-guided representations of the input. Across three relation extraction tasks, our method, ExpBERT, matches a BERT baseline but with 3–20x less labeled data and improves on the baseline by 3–10 F1 points with the same amount of labeled data.

GAN-BERT: Generative Adversarial Learning for Robust Text Classification with a Bunch of Labeled Examples [Website][PDF]
Danilo Croce, Giuseppe Castellucci, and Roberto Basili 20:00–21:00

Recent Transformer-based architectures, e.g., BERT, provide impressive results in many Natural Language Processing tasks. However, most of the adopted benchmarks are made of (sometimes hundreds of) thousands of examples. In many real scenarios, obtaining high- quality annotated data is expensive and time consuming; in contrast, unlabeled examples characterizing the target task can be, in general, easily collected. One promising method to enable semi-supervised learning has been proposed in image processing, based on Semi- Supervised Generative Adversarial Networks. In this paper, we propose GAN-BERT that ex- tends the fine-tuning of BERT-like architectures with unlabeled data in a generative adversarial setting. Experimental results show that the requirement for annotated examples can be drastically reduced (up to only 50-100 annotated examples), still obtaining good performances in several sentence classification tasks.

Generalizing Natural Language Analysis through Span-relation Representations [Website][PDF]
Zhengbao Jiang, Wei Xu, Jun Araki, and Graham Neubig 20:00–21:00

Natural language processing covers a wide variety of tasks predicting syntax, semantics, and information content, and usually each type of output is generated with specially designed architectures. In this paper, we provide the simple insight that a great variety of tasks can be represented in a single unified format consisting of labeling spans and relations between spans, thus a single task-independent model can be used across different tasks. We perform extensive experiments to test this insight on 10 disparate tasks spanning dependency parsing (syntax), semantic role labeling (semantics), relation extraction (information content), aspect based sentiment analysis (sentiment), and many others, achieving performance comparable to state-of-the-art specialized models. We further demonstrate benefits of multi-task learning, and also show that the proposed method makes it easy to analyze differences and similarities in how the model handles different tasks. Finally, we convert these datasets into a unified format to build a benchmark, which provides a holistic testbed for evaluating future models for generalized natural language analysis.

Learning to Contextually Aggregate Multi-Source Supervision for Sequence Labeling [Website][PDF]
Ouyu Lan, Xiao Huang, Bill Yuchen Lin, He Jiang, Liyuan Liu, and Xiang Ren 20:00–21:00

Sequence labeling is a fundamental task for a range of natural language processing problems. When used in practice, its performance is largely influenced by the annotation quality and quantity, and meanwhile, obtaining ground truth labels is often costly. In many cases, ground truth labels do not exist, but noisy annotations or annotations from dif-

ferent domains are accessible. In this paper, we propose a novel framework Consensus Network (ConNet) that can be trained on annotations from multiple sources (e.g., crowd annotation, cross-domain data). It learns individual representation for every source and dynamically aggregates source-specific knowledge by a context-aware attention module. Finally, it leads to a model reflecting the agreement (consensus) among multiple sources. We evaluate the proposed framework in two practical settings of multi-source learning: learning with crowd annotations and unsupervised cross-domain model adaptation. Extensive experimental results show that our model achieves significant improvements over existing methods in both settings. We also demonstrate that the method can apply to various tasks and cope with different encoders.

MixText: Linguistically-Informed Interpolation of Hidden Space for Semi-Supervised Text Classification [Website][PDF]
Jiaao Chen, Zichao Yang, and Diyi Yang 20:00–21:00

This paper presents MixText, a semi-supervised learning method for text classification, which uses our newly designed data augmentation method called TMix. TMix creates a large amount of augmented training samples by interpolating text in hidden space. Moreover, we leverage recent advances in data augmentation to guess low-entropy labels for unlabeled data, hence making them as easy to use as labeled data. By mixing labeled, unlabeled and augmented data, MixText significantly outperformed current pre-trained and fined-tuned models and other state-of-the-art semi-supervised learning methods on several text classification benchmarks. The improvement is especially prominent when supervision is extremely limited. We have publicly released our code at https://github.com/GT-SALT/MixText.

MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices [Website][PDF]
Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou 20:00–21:00

Natural Language Processing (NLP) has recently achieved great success by using huge pre-trained models with hundreds of millions of parameters. However, these models suffer from heavy model sizes and high latency such that they cannot be deployed to resource-limited mobile devices. In this paper, we propose MobileBERT for compressing and accelerating the popular BERT model. Like the original BERT, MobileBERT is task-agnostic, that is, it can be generically applied to various downstream NLP tasks via simple fine-tuning. Basically, MobileBERT is a thin version of BERT_LARGE, while equipped with bottleneck structures and a carefully designed balance between self-attentions and feed-forward networks. To train MobileBERT, we first train a specially designed teacher model, an inverted-bottleneck incorporated BERT_LARGE model. Then, we conduct knowledge transfer from this teacher to MobileBERT. Empirical studies show that MobileBERT is 4.3x smaller and 5.5x faster than BERT_BASE while achieving competitive results on well-known benchmarks. On the natural language inference tasks of GLUE, MobileBERT achieves a GLUE score of 77.7 (0.6 lower than BERT_BASE), and 62 ms latency on a Pixel 4 phone. On the SQuAD v1.1/v2.0 question answering task, MobileBERT achieves a dev F1 score of 90.0/79.2 (1.5/2.1 higher than BERT_BASE).

On Importance Sampling-Based Evaluation of Latent Language Models [Website][PDF]
Robert L Logan IV, Matt Gardner, and Sameer Singh 20:00–21:00

Language models that use additional latent structures (e.g., syntax trees, coreference chains, knowledge graph links) provide several advantages over traditional language models. However, likelihood-based evaluation of these models is often intractable as it requires marginalizing over the latent space. Existing works avoid this issue by using importance sampling. Although this approach has asymptotic guarantees, analysis is rarely conducted on the effect of decisions such as sample size and choice of proposal distribution on the reported estimates. In this paper, we carry out this analysis for three models: RNNG, EntityNLM, and KGLM. In addition, we elucidate subtle differences in how importance sampling is applied in these works that can have substantial effects on the final estimates, as well as provide theoretical results which reinforce the validity of this technique.

SMART: Robust and Efficient Fine-Tuning for Pre-trained Natural Language Models through Principled Regularized Optimization [Website][PDF]
Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao 20:00–21:00

Transfer learning has fundamentally changed the landscape of natural language processing (NLP). Many state-of-the-art models are first pre-trained on a large text corpus and then fine-tuned on downstream tasks. However, due to limited data resources from downstream tasks and the extremely high complexity of pre-trained models, aggressive fine-tuning often causes the fine-tuned model to overfit the training data of downstream tasks and fail to generalize to unseen data. To address such an issue in a principled manner, we propose a new learning framework for robust and efficient fine-tuning for pre-trained models to attain better generalization performance. The proposed framework contains two important ingredients: 1. Smoothness-inducing regularization, which effectively manages the complexity of the model; 2. Bregman proximal point optimization, which is an instance of trust-region methods and can prevent aggressive updating. Our experiments show that the proposed framework achieves new state-of-the-art performance on a number of NLP tasks including GLUE, SNLI, SciTail and ANLI. Moreover, it also outperforms the state-of-the-art T5 model, which is the largest pre-trained model containing 11 billion parameters, on GLUE.

Stolen Probability: A Structural Weakness of Neural Language Models [Website][PDF]
David Demeter, Gregory Kimmel, and Doug Downey 20:00–21:00

Neural Network Language Models (NNLMs) generate probability distributions by applying a softmax function to a distance metric formed by taking the dot product of a prediction vector with all word vectors in a high-dimensional embedding space. The dot-product distance metric forms part of the inductive bias of NNLMs. Although NNLMs optimize well with this inductive bias, we show that this results in a sub-optimal ordering of the embedding space that structurally impoverishes some words at the expense of others when assigning probability. We present numerical, theoretical and empirical analyses which show that words on the interior of the convex hull in the embedding space have their probability bounded by the probabilities of the words on the hull.

Taxonomy Construction of Unseen Domains via Graph-based Cross-Domain Knowledge Transfer [Website][PDF]
Chao Shang, Sarthak Dash, Md. Faisal Mahbub Chowdhury, Nandana Mihindukulasooriya, and Alfio Gliozzo 20:00–21:00

Extracting lexico-semantic relations as graph-structured taxonomies, also known as taxonomy construction, has been beneficial in a variety of NLP applications. Recently Graph Neural Network (GNN) has shown to be powerful in successfully tackling many tasks. However, there has been no attempt to exploit GNN to create taxonomies. In this paper, we propose Graph2Taxo, a GNN-based cross-domain transfer framework for the taxonomy construction task. Our main contribution is to learn the latent features of taxonomy construction from existing domains to guide the structure learning of an unseen domain. We also propose a novel method of directed acyclic graph (DAG) generation for taxonomy construction. Specifically, our proposed Graph2Taxo uses a noisy graph constructed from automatically extracted noisy hyponym hypernym candidate pairs, and a set of taxonomies for some known domains for training. The learned model is then used to generate taxonomy for a new unknown domain given a set of terms for that domain. Experiments on benchmark datasets from science and environment domains show that our approach attains significant improvements correspondingly over the state of the art.

To Pretrain or Not to Pretrain: Examining the Benefits of Pretrainng on Resource Rich Tasks [Website][PDF]
Sinong Wang, Madian Khabsa, and Hao Ma 20:00–21:00

Pretraining NLP models with variants of Masked Language Model (MLM) objectives has recently led to a significant improvements on many tasks. This paper examines the benefits of pretrained models as a function of the number of training samples used in the downstream task. On several text classification tasks, we show that as the number of training examples grow into the millions, the accuracy gap between finetuning BERT-based model and training vanilla LSTM from scratch narrows to within 1%. Our findings indicate that MLM-based models might reach a diminishing return point as the supervised data size increases significantly.

Why Overfitting Isn't Always Bad: Retrofitting Cross-Lingual Word Embeddings to Dictionaries [Website][PDF]
Mozhi Zhang, Yoshinari Fujinuma, Michael J. Paul, and Jordan Boyd-Graber 20:00–21:00

Cross-lingual word embeddings (CLWE) are often evaluated on bilingual lexicon induction (BLI). Recent CLWE methods use linear projections, which underfit the training dictionary, to generalize on BLI. However, underfitting can hinder generalization to other downstream tasks that rely on words from the training dictionary. We address this limitation by retrofitting CLWE to the training dictionary, which pulls training translation pairs closer in the embedding space and overfits the training dictionary. This simple post-processing step often improves accuracy on two downstream tasks, despite lowering BLI test accuracy. We also retrofit to both the training dictionary and a synthetic dictionary induced from CLWE, which sometimes generalizes even better on downstream tasks. Our results confirm the importance of fully exploiting training dictionary in downstream tasks and explains why BLI is a flawed CLWE evaluation.

XtremeDistil: Multi-stage Distillation for Massive Multilingual Models [Website][PDF]
Subhabrata Mukherjee and Ahmed Hassan Awadallah 20:00–21:00

Deep and large pre-trained language models are the state-of-the-art for various natural language processing tasks. However, the huge size of these models could be a deterrent to using them in practice. Some recent works use knowledge distillation to compress these huge models into shallow ones. In this work we study knowledge distillation with a focus on multilingual Named Entity Recognition (NER). In particular, we study several distillation strategies and propose a stage-wise optimization scheme leveraging teacher internal representations, that is agnostic of teacher architecture, and show that it outperforms strategies employed in prior works. Additionally, we investigate the role of several factors like the amount of unlabeled data, annotation resources, model architecture and inference latency to name a few. We show that our approach leads to massive compression of teacher models like mBERT by upto 35x in terms of parameters and 51x in terms of latency for batch inference while retaining 95% of its F1-score for NER over 41 languages.

Session 5A: Machine Translation-6

ENGINE: Energy-Based Inference Networks for Non-Autoregressive Machine Translation [Website][PDF]
Lifu Tu, Richard Yuanzhe Pang, Sam Wiseman, and Kevin Gimpel 20:00–21:00

We propose to train a non-autoregressive machine translation model to minimize the energy defined by a pretrained autoregressive model. In particular, we view our non-autoregressive translation system as an inference network (Tu and Gimpel, 2018) trained to minimize the autoregressive teacher energy. This contrasts with the popular approach of training a non-autoregressive model on a distilled corpus consisting of the beam-searched outputs of such a teacher model. Our approach, which we call ENGINE (ENerGy-based Inference NEtworks), achieves state-of-the-art non-autoregressive results on the IWSLT 2014 DE-EN and WMT 2016 RO-EN datasets, approaching the performance of autoregressive models.

Improving Non-autoregressive Neural Machine Translation with Monolingual Data [Website][PDF]
Jiawei Zhou and Phillip Keung 20:00–21:00

Non-autoregressive (NAR) neural machine translation is usually done via knowledge distillation from an autoregressive (AR) model. Under this framework, we leverage large monolingual corpora to improve the NAR model's performance, with the goal of transferring the AR model's generalization ability while preventing overfitting. On top of a strong NAR baseline, our experimental results on the WMT14 En-De and WMT16 En-Ro news translation tasks confirm that monolingual data augmentation consistently improves the performance of the NAR model to approach the teacher AR model's performance, yields comparable or better results than the best non-iterative NAR methods in the literature and helps reduce overfitting in the training process.

Leveraging Monolingual Data with Self-Supervision for Multilingual Neural Machine Translation [Website][PDF]
Aditya Siddhant, Ankur Bapna, Yuan Cao, Orhan Firat, Mia Chen, Sneha Kudugunta, Naveen Arivazhagan, and Yonghui Wu 20:00–21:00

Over the last few years two promising research directions in low-resource neural machine translation (NMT) have emerged. The first focuses on utilizing high-resource languages to improve the quality of low-resource languages via multilingual NMT. The second direction employs monolingual data with self-supervision to pre-train translation models, followed by fine-tuning on small amounts of supervised data. In this work, we join these two lines of research and demonstrate the efficacy of monolingual data with self-supervision in multilingual NMT. We offer three major results: (i) Using monolingual data significantly boosts the translation quality of low-resource languages in multilingual models. (ii) Self-supervision improves zero-shot translation quality in multilingual models. (iii) Leveraging monolingual data with self-supervision provides a viable path towards adding new languages to multilingual models, getting up to 33 BLEU on ro-en translation without any parallel data or back-translation.

Location Attention for Extrapolation to Longer Sequences [Website][PDF]
Yann Dubois, Gautier Dagan, Dieuwke Hupkes, and Elia Bruni 20:00–21:00

Neural networks are surprisingly good at interpolating and perform remarkably well when the training set examples resemble those in the test set. However, they are often unable to extrapolate patterns beyond the seen data, even when the abstractions required for such patterns are simple. In this paper, we first review the notion of extrapolation, why it is important and how one could hope to tackle it. We then focus on a specific type of extrapolation which is especially useful for natural language processing: generalization to sequences that are longer than the training ones. We hypothesize that models with a separate content- and location-based attention are more likely to extrapolate than those with common attention mechanisms. We empirically support our claim for recurrent seq2seq models with our proposed attention on variants of the Lookup Table task. This sheds light on some striking failures of neural models for sequences and on possible methods to approaching such issues.

On The Evaluation of Machine Translation SystemsTrained With Back-Translation [Website][PDF]
Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, and Michael Auli 20:00–21:00

Back-translation is a widely used data augmentation technique which leverages target monolingual data. However, its effectiveness has been challenged since automatic metrics such as BLEU only show significant improvements for test examples where the source itself is a translation, or translationese. This is believed to be due to translationese inputs better matching the back-translated training data. In this work, we show that this conjecture is not empirically supported and that back-translation improves translation quality of both naturally occurring text as well as translationese according to professional human translators. We provide empirical evidence to support the view that back-translation is preferred by humans because it produces more fluent outputs. BLEU cannot capture human preferences because references are translationese when source sentences are natural text. We recommend complementing BLEU with a language model score to measure fluency.

Opportunistic Decoding with Timely Correction for Simultaneous Translation [Website][PDF]
Renjie Zheng, Mingbo Ma, Baigong Zheng, Kaibo Liu, and Liang Huang 20:00–21:00

Simultaneous translation has many important application scenarios and attracts much attention from both academia and industry recently. Most existing frameworks, however, have difficulties in balancing between the translation quality and latency, i.e., the decoding policy is usually either too aggressive or too conservative. We propose an opportunistic decoding technique with timely correction ability, which always (over-)generates a certain mount of extra words at each step to keep the audience on track with the latest information. At the same time, it also corrects, in a timely fash-

ion, the mistakes in the former overgenerated words when observing more source context to ensure high translation quality. Experiments show our technique achieves substantial reduction in latency and up to +3.1 increase in BLEU, with revision rate under 8% in Chinese-to-English and English-to-Chinese translation.

Simultaneous Translation Policies: From Fixed to Adaptive [Website] [PDF]
Baigong Zheng, Kaibo Liu, Renjie Zheng, Mingbo Ma, Hairong Liu, and Liang Huang 20:00–21:00

Adaptive policies are better than fixed policies for simultaneous translation, since they can flexibly balance the trade-off between translation quality and latency based on the current context information. But previous methods on obtaining adaptive policies either rely on complicated training process, or underperform simple fixed policies. We design an algorithm to achieve adaptive policies via a simple heuristic composition of a set of fixed policies. Experiments on Chinese -> English and German -> English show that our adaptive policies can outperform fixed ones by up to 4 BLEU points for the same latency, and more surprisingly, it even surpasses the BLEU score of full-sentence translation in the greedy mode (and very close to beam mode), but with much lower latency.

Session 5A Semantics: Textual Inference and Other Areas of Semantics-2

Can We Predict New Facts with Open Knowledge Graph Embeddings? A Benchmark for Open Link Prediction [Website][PDF]
Samuel Broscheit, Kiril Gashteovski, Yanjie Wang, and Rainer Gemulla 20:00–21:00

Open Information Extraction systems extract ("subject text", "relation text", "object text") triples from raw text. Some triples are textual versions of facts, i.e., non-canonicalized mentions of entities and relations. In this paper, we investigate whether it is possible to infer new facts directly from the open knowledge graph without any canonicalization or any supervision from curated knowledge. For this purpose, we propose the open link prediction task,i.e., predicting test facts by completing ("subject text", "relation text", ?) questions. An evaluation in such a setup raises the question if a correct prediction is actually a new fact that was induced by reasoning over the open knowledge graph or if it can be trivially explained. For example, facts can appear in different paraphrased textual variants, which can lead to test leakage. To this end, we propose an evaluation protocol and a methodology for creating the open link prediction benchmark OlpBench. We performed experiments with a prototypical knowledge graph embedding model for open-link prediction. While the task is very challenging, our results suggests that it is possible to predict genuinely new facts, which can not be trivially explained.

[TACL] Decomposing Generalization: Models of Generic, Habitual and Episodic Statements [Website][PDF]
Venkata Subrahmanyan Govindarajan, Benjamin Van Durme, and Aaron Steven White 20:00–21:00

We present a novel semantic framework for modeling linguistic expressions of generalization—generic, habitual, and episodic statements—as combinations of simple, real-valued referential properties of predicates and their arguments. We use this framework to construct a dataset covering the entirety of the Universal Dependencies English Web Treebank. We use this dataset to probe the efficacy of type-level and token-level information—including hand-engineered features and static (GloVe) and contextual (ELMo) word embeddings—for predicting expressions of generalization.

INFOTABS: Inference on Tables as Semi-structured Data [Website][PDF]
Vivek Gupta, Maitrey Mehta, Pegah Nokhiz, and Vivek Srikumar 20:00–21:00

In this paper, we observe that semi-structured tabulated text is ubiquitous; understanding them requires not only comprehending the meaning of text fragments, but also implicit relationships between them. We argue that such data can prove as a testing ground for understanding how we reason about information. To study this, we introduce a new dataset called INFOTABS, comprising of human-written textual hypotheses based on premises that are tables extracted from Wikipedia info-boxes. Our analysis shows that the semi-structured, multi-domain and heterogeneous nature of the premises admits complex, multi-faceted reasoning. Experiments reveal that, while human annotators agree on the relationships between a table-hypothesis pair, several standard modeling strategies are unsuccessful at the task, suggesting that reasoning about tables can pose a difficult modeling challenge.

[TACL] Inherent Disagreements in Human Textual Inferences [Website][PDF]
Ellie Pavlick and Tom Kwiatkowski 20:00–21:00

We analyze human's disagreements about the validity of natural language inferences. We show that, very often, disagreements are not dismissible as annotation "noise", but rather persist as we collect more ratings and as we vary the amount of context provided to raters. We further show that the type of uncertainty captured by current state-of-the-art models for natural language inference is not reflective of the type of uncertainty present in human disagreements. We discuss implications of our results in relation to the recognizing textual entailment (RTE)/natural language inference (NLI) task. We argue for a refined evaluation objective which requires models to explicitly capture the full distribution of plausible human judgments.

Interactive Machine Comprehension with Information Seeking Agents [Website][PDF]
Xingdi Yuan, Jie Fu, Marc-Alexandre Côté, Yi Tay, Chris Pal, and Adam Trischler 20:00–21:00

Existing machine reading comprehension (MRC) models do not scale effectively to real-world applications like web-level information retrieval and question answering (QA). We argue that this stems from the nature of MRC datasets: most of these are static environments wherein the supporting documents and all necessary information are fully observed. In this paper, we propose a simple method that reframes existing MRC datasets as interactive, partially observable environments. Specifically, we "occlude" the majority of a document's text and add context-sensitive commands that reveal "glimpses" of the hidden text to a model. We repurpose SQuAD and NewsQA as an initial case study, and then show how the interactive corpora can be used to train a model that seeks relevant information through sequential decision making. We believe that this setting can contribute in scaling models to web-level QA scenarios.

Syntactic Data Augmentation Increases Robustness to Inference Heuristics [Website][PDF]
Junghyun Min, R. Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen 20:00–21:00

Pretrained neural models such as BERT, when fine-tuned to perform natural language inference (NLI), often show high accuracy on standard datasets, but display a surprising lack of sensitivity to word order on controlled challenge sets. We hypothesize that this issue is not primarily caused by the pretrained model's limitations, but rather by the paucity of crowdsourced NLI examples that might convey the importance of syntactic structure at the fine-tuning stage. We explore several methods to augment standard training sets with syntactically informative examples, generated by applying syntactic transformations to sentences from the MNLI corpus. The best-performing augmentation method, subject/object inversion, improved BERT's accuracy on controlled examples that diagnose sensitivity to word order from 0.28 to 0.73, without affecting performance on the MNLI test set. This improvement generalized be-

yond the particular construction used for data augmentation, suggesting that augmentation causes BERT to recruit abstract syntactic representations.

Session 5A: Speech and Multimodality-2

[TACL] Acoustic-Prosodic and Lexical Cues to Deception and Trust: Deciphering How People Detect Lies [Website][PDF]
Xi (Leslie) Chen, Sarah Ita Levitan, Michelle Levine, Marko Mandic, and Julia Hirschberg 20:00–21:00

Humans rarely perform better than chance at lie detection. To better understand human perception of deception, we created a game framework, LieCatcher, to collect ratings of perceived deception using a large corpus of deceptive and truthful interviews. We analyzed the acoustic-prosodic and linguistic characteristics of language trusted and mistrusted by raters and compared these to characteristics of actual truthful and deceptive language to understand how perception aligns with reality. With this data we built classifiers to automatically distinguish trusted from mistrusted speech, achieving an F1 of 66.1%. We next evaluated whether the strategies raters said they used to discriminate between truthful and deceptive responses were in fact useful. Our results show that, while several prosodic and lexical features were consistently perceived as trustworthy, they were not reliable cues. Also, the strategies that judges reported using in deception detection were not helpful for the task. Our work sheds light on the nature of trusted language and provides insight into the challenging problem of human deception detection.

Improved Speech Representations with Multi-Target Autoregressive Predictive Coding [Website][PDF]
Yu-An Chung and James Glass 20:00–21:00

Training objectives based on predictive coding have recently been shown to be very effective at learning meaningful representations from unlabeled speech. One example is Autoregressive Predictive Coding (Chung et al., 2019), which trains an autoregressive RNN to generate an unseen future frame given a context such as recent past frames. The basic hypothesis of these approaches is that hidden states that can accurately predict future frames are a useful representation for many downstream tasks. In this paper we extend this hypothesis and aim to enrich the information encoded in the hidden states by training the model to make more accurate future predictions. We propose an auxiliary objective that serves as a regularization to improve generalization of the future frame prediction task. Experimental results on phonetic classification, speech recognition, and speech translation not only support the hypothesis, but also demonstrate the effectiveness of our approach in learning representations that contain richer phonetic content.

Integrating Multimodal Information in Large Pretrained Transformers [Website][PDF]
Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque 20:00–21:00

Recent Transformer-based contextual word representations, including BERT and XLNet, have shown state-of-the-art performance in multiple disciplines within NLP. Fine-tuning the trained contextual models on task-specific datasets has been the key to achieving superior performance downstream. While fine-tuning these pre-trained models is straightforward for lexical applications (applications with only language modality), it is not trivial for multimodal language (a growing area in NLP focused on modeling face-to-face communication). More specifically, this is due to the fact that pre-trained models don't have the necessary components to accept two extra modalities of vision and acoustic. In this paper, we proposed an attachment to BERT and XLNet called Multimodal Adaptation Gate (MAG). MAG allows BERT and XLNet to accept multimodal nonverbal data during fine-tuning. It does so by generating a shift to internal representation of BERT and XLNet; a shift that is conditioned on the visual and acoustic modalities. In our experiments, we study the commonly used CMU-MOSI and CMU-MOSEI datasets for multimodal sentiment analysis. Fine-tuning MAG-BERT and MAG-XLNet significantly boosts the sentiment analysis performance over previous baselines as well as language-only fine-tuning of BERT and XLNet. On the CMU-MOSI dataset, MAG-XLNet achieves human-level multimodal sentiment analysis performance for the first time in the NLP community.

MultiQT: Multimodal learning for real-time question tracking in speech [Website][PDF]
Jakob D. Havtorn, Jan Latko, Joakim Edin, Lars Maaløe, Lasse Borgholt, Lorenzo Belgrano, Nicolai Jacobsen, Regitze Sdun, and Željko Agić 20:00–21:00

We address a challenging and practical task of labeling questions in speech in real time during telephone calls to emergency medical services in English, which embeds within a broader decision support system for emergency call-takers. We propose a novel multimodal approach to real-time sequence labeling in speech. Our model treats speech and its own textual representation as two separate modalities or views, as it jointly learns from streamed audio and its noisy transcription into text via automatic speech recognition. Our results show significant gains of jointly learning from the two modalities when compared to text or audio only, under adverse noise and limited volume of training data. The results generalize to medical symptoms detection where we observe a similar pattern of improvements with multimodal learning.

Multimodal and Multiresolution Speech Recognition with Transformers [Website][PDF]
Georgios Paraskevopoulos, Srinivas Parthasarathy, Aparna Khare, and Shiva Sundaram 20:00–21:00

This paper presents an audio visual automatic speech recognition (AV-ASR) system using a Transformer-based architecture. We particularly focus on the scene context provided by the visual information, to ground the ASR. We extract representations for audio features in the encoder layers of the transformer and fuse video features using an additional crossmodal multihead attention layer. Additionally, we incorporate a multitask training criterion for multiresolution ASR, where we train the model to generate both character and subword level transcriptions. Experimental results on the How2 dataset, indicate that multiresolution training can speed up convergence by around 50% and relatively improves word error rate (WER) performance by upto 18% over subword prediction models. Further, incorporating visual information improves performance with relative gains upto 3.76% over audio only models. Our results are comparable to state-of-the-art Listen, Attend and Spell-based architectures.

Phone Features Improve Speech Translation [Website][PDF]
Elizabeth Salesky and Alan W Black 20:00–21:00

End-to-end models for speech translation (ST) more tightly couple speech recognition (ASR) and machine translation (MT) than a traditional cascade of separate ASR and MT models, with simpler model architectures and the potential for reduced error propagation. Their performance is often assumed to be superior, though in many conditions this is not yet the case. We compare cascaded and end-to-end models across high, medium, and low-resource conditions, and show that cascades remain stronger baselines. Further, we introduce two methods to incorporate phone features into ST models. We show that these features improve both architectures, closing the gap between end-to-end models and cascades, and outperforming previous academic work – by up to 9 BLEU on our low-resource setting.

Session 5A: Theory and Formalism in NLP (Linguistic and Mathematical)-3

A Formal Hierarchy of RNN Architectures [Website][PDF]
William Merrill, Gail Weiss, Yoav Goldberg, Roy Schwartz, Noah A. Smith, and Eran Yahav 20:00–21:00

We develop a formal hierarchy of the expressive capacity of RNN architectures. The hierarchy is based on two formal properties: space complexity, which measures the RNN's memory, and rational recurrence, defined as whether the recurrent update can be described by a weighted finite-state machine. We place several RNN variants within this hierarchy. For example, we prove the LSTM is not rational, which formally separates it from the related QRNN (Bradbury et al., 2016). We also show how these models' expressive capacity is expanded by stacking multiple layers or composing them with different pooling functions. Our results build on the theory of "saturated" RNNs (Merrill, 2019). While formally extending these findings to unsaturated RNNs is left to future work, we hypothesize that the practical learnable capacity of unsaturated RNNs obeys a similar hierarchy. We provide empirical results to support this conjecture. Experimental findings from training unsaturated networks on formal languages support this conjecture.

Emergence of Syntax Needs Minimal Supervision [Website][PDF]
Raphaël Bailly and Kata Gábor 20:00–21:00

This paper is a theoretical contribution to the debate on the learnability of syntax from a corpus without explicit syntax-specific guidance. Our approach originates in the observable structure of a corpus, which we use to define and isolate grammaticality (syntactic information) and meaning/pragmatics information. We describe the formal characteristics of an autonomous syntax and show that it becomes possible to search for syntax-based lexical categories with a simple optimization process, without any prior hypothesis on the form of the model.

Demo Session 5B

Time: *20:45–21:30*

Multilingual Universal Sentence Encoder for Semantic Retrieval [Website][PDF]
Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil

We present easy-to-use retrieval focused multilingual sentence embedding models, made available on TensorFlow Hub. The models embed text from 16 languages into a shared semantic space using a multi-task trained dual-encoder that learns tied cross-lingual representations via translation bridge tasks (Chidambaram et al., 2018). The models achieve a new state-of-the-art in performance on monolingual and cross-lingual semantic retrieval (SR). Competitive performance is obtained on the related tasks of translation pair bitext retrieval (BR) and retrieval question answering (ReQA). On transfer learning tasks, our multilingual embeddings approach, and in some cases exceed, the performance of English only sentence embeddings.

SyntaxGym: An Online Platform for Targeted Evaluation of Language Models [Website][PDF]
Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy

Targeted syntactic evaluations have yielded insights into the generalizations learned by neural network language models. However, this line of research requires an uncommon confluence of skills: both the theoretical knowledge needed to design controlled psycholinguistic experiments, and the technical proficiency needed to train and deploy large-scale language models. We present SyntaxGym, an online platform designed to make targeted evaluations accessible to both experts in NLP and linguistics, reproducible across computing environments, and standardized following the norms of psycholinguistic experimental design. This paper releases two tools of independent value for the computational linguistics community: 1. A website, syntaxgym.org, which centralizes the process of targeted syntactic evaluation and provides easy tools for analysis and visualization; 2. Two command-line tools, ‘syntaxgym‘ and ‘lm-zoo‘, which allow any user to reproduce targeted syntactic evaluations and general language model inference on their own machine.

Session 5B Overview – Monday, July 6, 2020 21:00–22:00

Track A *Cognitive Modeling and Psycholinguistics-5* Abstracts	A Systematic Assessment of Syntactic Generalization in Neural Language Models *Hu, Gauthier, Qian, Wilcox, and Levy* [Website][PDF]	A Tale of Two Perplexities: Sensitivity of Neural Language Models to Lexical Retrieval Deficits in Dementia of the Alzheimer's Type *Cohen and Pakhomov* [Website][PDF]	Learning to Understand Child-directed and Adult-directed Speech *Gelderloos, Chrupała, and Alishahi* [Website][PDF]	Overestimation of Syntactic Representation in Neural Language Models *Kodner and Gupta* [Website][PDF]	Probing Linguistic Systematicity *Goodwin, Sinha, and O'Donnell* [Website][PDF]
	Recollection versus Imagination: Exploring Human Memory and Cognition via Neural Language Models *Sap, horvitz, Choi, Smith, and Pennebaker* [Website][PDF]	Recurrent Neural Network Language Models Always Learn English-Like Relative Clause Attachment *Davis and Schijndel* [Website][PDF]	Speakers enhance contextually confusable words *Meinhardt, Bakovic, and Bergen* [Website][PDF]	Suspense in Short Stories is Predicted By Uncertainty Reduction over Neural Story Representation *Wilmot and Keller* [Website][PDF]	What determines the order of adjectives in English? Comparing efficiency-based theories using dependency treebanks *Futrell, Dyer, and Scontras* [Website][PDF]
Track B *Dialogue and Interactive Systems-10* Abstracts	"None of the Above": Measure Uncertainty in Dialog Response Retrieval *Feng, Mehri, Eskenazi, and Zhao* [Website][PDF]	Can You Put it All Together: Evaluating Conversational Agents' Ability to Blend Skills *Smith, Williamson, Shuster, Weston, and Boureau* [Website][PDF]	Conversational Graph Grounded Policy Learning for Open-Domain Conversation Generation *Xu, Wang, Niu, Wu, Che, and Liu* [Website][PDF]	Grounding Conversations with Improvised Dialogues *Cho and May* [Website][PDF]	Image-Chat: Engaging Grounded Conversations *Shuster, Humeau, Bordes, and Weston* [Website][PDF]
	Large Scale Multi-Actor Generative Dialog Modeling *Boyd, Puri, Shoeybi, Patwary, and Catanzaro* [Website][PDF]	Neural Generation of Dialogue Response Timings *Roddy and Harte* [Website][PDF]	[CL] The Design and Implementation of XiaoIce, an Empathetic Social Chatbot *Zhou, Gao, Li, and Shum* [Website][PDF]	The Dialogue Dodecathlon: Open-Domain Knowledge and Image Grounded Conversational Agents *Shuster, JU, Roller, Dinan, Boureau, and Weston* [Website][PDF]	
Track C *Language Grounding to Vision, Robotics and Beyond-2* Abstracts	BabyWalk: Going Farther in Vision-and-Language Navigation by Taking Baby Steps *Zhu, Hu, Chen, Deng, Jain, Ie, and Sha* [Website][PDF]	Cross-media Structured Common Space for Multimedia Event Extraction *Li, Zareian, Zeng, Whitehead, Lu, Ji, and Chang* [Website][PDF]	Learning to Segment Actions from Observation and Narration *Fried, Alayrac, Blunsom, Dyer, Clark, and Nematzadeh* [Website][PDF]	Learning to execute instructions in a Minecraft dialogue *Jayannavar, Narayan-Chen, and Hockenmaier* [Website][PDF]	MART: Memory-Augmented Recurrent Transformer for Coherent Video Paragraph Captioning *Lei, Wang, Shen, Yu, Berg, and Bansal* [Website][PDF]
	What is Learned in Visually Grounded Neural Syntax Acquisition *Kojima, Averbuch-Elor, Rush, and Artzi* [Website][PDF]				

Track D *Machine Learning for NLP-4* Abstracts	A Batch Normalized Inference Network Keeps the KL Vanishing Away *Zhu, Bi, Liu, Ma, Li, and Wu* [Website][PDF]	Contextual Embeddings: When Are They Worth It? *Arora, May, Zhang, and Ré* [Website][PDF]	[TACL] Efficient Contextual Representation Learning With Continuous Outputs *Li, Chen, Hsieh, and Chang* [Website][PDF]	Interactive Classification by Asking Informative Questions *Yu, Chen, Wang, Lei, and Artzi* [Website][PDF]	Knowledge Graph Embedding Compression *Sachan* [Website][PDF]
	Low Resource Sequence Tagging using Sentence Reconstruction *Perl, Chaudhury, and Giryes* [Website][PDF]	Masked Language Model Scoring *Salazar, Liang, Nguyen, and Kirchhoff* [Website][PDF]	Orthogonal Relation Transforms with Graph Context Modeling for Knowledge Graph Embedding *Tang, Huang, Wang, He, and Zhou* [Website][PDF]	[TACL] Perturbation Based Learning for Structured NLP Tasks with Application to Dependency Parsing *Doitch, Yazdi, Hazan, and Reichart* [Website][PDF]	Posterior Calibrated Training on Sentence Classification Tasks *Jung, Kang, Cheng, Mentch, and Schaaf* [Website][PDF]
	Posterior Control of Blackbox Generation *Li and Rush* [Website][PDF]	Pretrained Transformers Improve Out-of-Distribution Robustness *Hendrycks, Liu, Wallace, Dziedzic, Krishnan, and Song* [Website][PDF]	Robust Encodings: A Framework for Combating Adversarial Typos *Jones, Jia, Raghunathan, and Liang* [Website][PDF]	Showing Your Work Doesn't Always Work *Tang, Lee, Xin, Liu, Yu, and Lin* [Website][PDF]	Span Selection Pre-training for Question Answering *Glass, Gliozzo, Chakravarti, Ferritto, Pan, Bhargav, Garg, and Sil* [Website][PDF]
	Topological Sort for Sentence Ordering *Prabhumoye, Salakhutdinov, and Black* [Website][PDF]	Weight Poisoning Attacks on Pretrained Models *Kurita, Michel, and Neubig* [Website][PDF]	schuBERT: Optimizing Elements of BERT *Khetan and Karnin* [Website][PDF]		
Track E *NLP Applications-4* Abstracts	"The Boating Store Had Its Best Sail Ever": Pronunciation-attentive Contextualized Pun Recognition *Zhou, Jiang, Zhao, Chang, and Wang* [Website][PDF]	A Girl Has A Name: Detecting Authorship Obfuscation *Mahmood, Shafiq, and Srinivasan* [Website][PDF]	DeeBERT: Dynamic Early Exiting for Accelerating BERT Inference *Xin, Tang, Lee, Yu, and Lin* [Website][PDF]	Efficient Strategies for Hierarchical Text Classification: External Knowledge and Auxiliary Tasks *Rivas Rojas, Bustamante, Onceway, and Sobrevilla Cabezudo* [Website][PDF]	Investigating the effect of auxiliary objectives for the automated grading of learner English speech transcriptions *Craighead, Caines, Buttery, and Yannakoudakis* [Website][PDF]
	SPECTER: Document-level Representation Learning using Citation-informed Transformers *Cohan, Feldman, Beltagy, Downey, and Weld* [Website][PDF]	Semantic Scaffolds for Pseudocode-to-Code Generation *Zhong, Stern, and Klein* [Website][PDF]			

Track F *Lexical-3* Abstracts	Breaking Through the 80% Glass Ceiling: Raising the State of the Art in Word Sense Disambiguation by Incorporating Knowledge Graph Information *Bevilacqua and Navigli* [Website][PDF]	Glyph2Vec: Learning Chinese Out-of-Vocabulary Word Embedding from Glyphs *Chen, YU, and Lin* [Website][PDF]	[TACL] Learning Lexical Subspaces in a Distributional Vector Space *Arora, Chakraborty, and Cheung* [Website][PDF]	Multidirectional Associative Optimization of Function-Specific Word Representations *Gerz, Vulić, Rei, Reichart, and Korhonen* [Website][PDF]	Predicting Degrees of Technicality in Automatic Terminology Extraction *Hätty, Schlechtweg, Dorna, and Schulte im Walde* [Website][PDF]
	Verbal Multiword Expressions for Identification of Metaphor *Rohanian, Rei, Taslimipoor, and Ha* [Website][PDF]				

Session 5B Details

Session 5B: Cognitive Modeling and Psycholinguistics-5

A Systematic Assessment of Syntactic Generalization in Neural Language Models [Website][PDF]
Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy 21:00–22:00

While state-of-the-art neural network models continue to achieve lower perplexity scores on language modeling benchmarks, it remains unknown whether optimizing for broad-coverage predictive performance leads to human-like syntactic knowledge. Furthermore, existing work has not provided a clear picture about the model properties required to produce proper syntactic generalizations. We present a systematic evaluation of the syntactic knowledge of neural language models, testing 20 combinations of model types and data sizes on a set of 34 English-language syntactic test suites. We find substantial differences in syntactic generalization performance by model architecture, with sequential models underperforming other architectures. Factorially manipulating model architecture and training dataset size (1M-40M words), we find that variability in syntactic generalization performance is substantially greater by architecture than by dataset size for the corpora tested in our experiments. Our results also reveal a dissociation between perplexity and syntactic generalization performance.

A Tale of Two Perplexities: Sensitivity of Neural Language Models to Lexical Retrieval Deficits in Dementia of the Alzheimer's Type [Website][PDF]
Trevor Cohen and Serguei Pakhomov 21:00–22:00

In recent years there has been a burgeoning interest in the use of computational methods to distinguish between elicited speech samples produced by patients with dementia, and those from healthy controls. The difference between perplexity estimates from two neural language models (LMs) - one trained on transcripts of speech produced by healthy participants and one trained on those with dementia - as a single feature for diagnostic classification of unseen transcripts has been shown to produce state-of-the-art performance. However, little is known about why this approach is effective, and on account of the lack of case/control matching in the most widely-used evaluation set of transcripts (DementiaBank), it is unclear if these approaches are truly diagnostic, or are sensitive to other variables. In this paper, we interrogate neural LMs trained on participants with and without dementia by using synthetic narratives previously developed to simulate progressive semantic dementia by manipulating lexical frequency. We find that perplexity of neural LMs is strongly and differentially associated with lexical frequency, and that using a mixture model resulting from interpolating control and dementia LMs improves upon the current state-of-the-art for models trained on transcript text exclusively.

Learning to Understand Child-directed and Adult-directed Speech [Website][PDF]
Lieke Gelderloos, Grzegorz Chrupała, and Afra Alishahi 21:00–22:00

Speech directed to children differs from adult-directed speech in linguistic aspects such as repetition, word choice, and sentence length, as well as in aspects of the speech signal itself, such as prosodic and phonemic variation. Human language acquisition research indicates that child-directed speech helps language learners. This study explores the effect of child-directed speech when learning to extract semantic information from speech directly. We compare the task performance of models trained on adult-directed speech (ADS) and child-directed speech (CDS). We find indications that CDS helps in the initial stages of learning, but eventually, models trained on ADS reach comparable task performance, and generalize better. The results suggest that this is at least partially due to linguistic rather than acoustic properties of the two registers, as we see the same pattern when looking at models trained on acoustically comparable synthetic speech.

Overestimation of Syntactic Representation in Neural Language Models [Website][PDF]
Jordan Kodner and Nitish Gupta 21:00–22:00

With the advent of powerful neural language models over the last few years, research attention has increasingly focused on what aspects of language they represent that make them so successful. Several testing methodologies have been developed to probe models' syntactic representations. One popular method for determining a model's ability to induce syntactic structure trains a model on strings generated according to a template then tests the model's ability to distinguish such strings from superficially similar ones with different syntax. We illustrate a fundamental problem with this approach by reproducing positive results from a recent paper with two non-syntactic baseline language models: an n-gram model and an LSTM model trained on scrambled inputs.

Probing Linguistic Systematicity [Website][PDF]
Emily Goodwin, Koustuv Sinha, and Timothy J. O'Donnell 21:00–22:00

Recently, there has been much interest in the question of whether deep natural language understanding (NLU) models exhibit systematicity, generalizing such that units like words make consistent contributions to the meaning of the sentences in which they appear. There is accumulating evidence that neural models do not learn systematically. We examine the notion of systematicity from a linguistic perspective, defining a set of probing tasks and a set of metrics to measure systematic behaviour. We also identify ways in which network architectures can generalize non-systematically, and discuss why such forms of generalization may be unsatisfying. As a case study, we perform a series of experiments in the setting of natural language inference (NLI). We provide evidence that current state-of-the-art NLU systems do not generalize systematically, despite overall high performance.

Recollection versus Imagination: Exploring Human Memory and Cognition via Neural Language Models [Website][PDF]
Maarten Sap, Eric horvitz, Yejin Choi, Noah A. Smith, and James Pennebaker 21:00–22:00

We investigate the use of NLP as a measure of the cognitive processes involved in storytelling, contrasting imagination and recollection of events. To facilitate this, we collect and release Hippocorpus, a dataset of 7,000 stories about imagined and recalled events. We introduce a measure of narrative flow and use this to examine the narratives for imagined and recalled events. Additionally, we measure the differential recruitment of knowledge attributed to semantic memory versus episodic memory (Tulving, 1972) for imagined and recalled storytelling by comparing the frequency of descriptions of general commonsense events with more specific realis events. Our analyses show that imagined stories have a substantially more linear narrative flow, compared to recalled stories in which adjacent sentences are more disconnected. In addition, while recalled stories rely more on autobiographical events based on episodic memory, imagined stories express more commonsense knowledge based on semantic memory. Finally, our measures reveal the effect of narrativization of memories in stories (e.g., stories about frequently recalled memories flow more linearly; Bartlett, 1932). Our findings highlight the potential of using NLP tools to study the traces of human cognition in language.

Recurrent Neural Network Language Models Always Learn English-Like Relative Clause Attachment [Website][PDF]
Forrest Davis and Marten van Schijndel 21:00–22:00

A standard approach to evaluating language models analyzes how models assign probabilities to valid versus invalid syntactic constructions (i.e. is a grammatical sentence more probable than an ungrammatical sentence). Our work uses ambiguous relative clause attachment to extend such evaluations to cases of multiple simultaneous valid interpretations, where stark grammaticality differences are absent. We compare model performance in English and Spanish to show that non-linguistic biases in RNN LMs advantageously overlap with syntactic structure in English but not Spanish. Thus, English models may appear to acquire human-like syntactic preferences, while models trained on Spanish fail to acquire comparable human-like preferences. We conclude by relating these results to broader concerns about the relationship between comprehension (i.e. typical language model use cases) and production (which generates the training data for language models), suggesting that necessary linguistic biases are not present in the training signal at all.

Speakers enhance contextually confusable words [Website][PDF]
Eric Meinhardt, Eric Bakovic, and Leon Bergen 21:00–22:00

Recent work has found evidence that natural languages are shaped by pressures for efficient communication — e.g. the more contextually predictable a word is, the fewer speech sounds or syllables it has (Piantadosi et al. 2011). Research on the degree to which speech and language are shaped by pressures for effective communication — robustness in the face of noise and uncertainty — has been more equivocal. We develop a measure of contextual confusability during word recognition based on psychoacoustic data. Applying this measure to naturalistic speech corpora, we find evidence suggesting that speakers alter their productions to make contextually more confusable words easier to understand.

Suspense in Short Stories is Predicted By Uncertainty Reduction over Neural Story Representation [Website][PDF]
David Wilmot and Frank Keller 21:00–22:00

Suspense is a crucial ingredient of narrative fiction, engaging readers and making stories compelling. While there is a vast theoretical literature on suspense, it is computationally not well understood. We compare two ways for modelling suspense: surprise, a backward-looking measure of how unexpected the current state is given the story so far; and uncertainty reduction, a forward-looking measure of how unexpected the continuation of the story is. Both can be computed either directly over story representations or over their probability distributions. We propose a hierarchical language model that encodes stories and computes surprise and uncertainty reduction. Evaluating against short stories annotated with human suspense judgements, we find that uncertainty reduction over representations is the best predictor, resulting in near human accuracy. We also show that uncertainty reduction can be used to predict suspenseful events in movie synopses.

What determines the order of adjectives in English? Comparing efficiency-based theories using dependency treebanks [Website][PDF]
Richard Futrell, William Dyer, and Greg Scontras 21:00–22:00

We take up the scientific question of what determines the preferred order of adjectives in English, in phrases such as big blue box where multiple adjectives modify a following noun. We implement and test four quantitative theories, all of which are theoretically motivated in terms of efficiency in human language production and comprehension. The four theories we test are subjectivity (Scontras et al., 2017), information locality (Futrell, 2019), integration cost (Dyer, 2017), and information gain, which we introduce. We evaluate theories based on their ability to predict orders of unseen adjectives in hand-parsed and automatically-parsed dependency treebanks. We find that subjectivity, information locality, and information gain are all strong predictors, with some evidence for a two-factor account, where subjectivity and information gain reflect a factor involving semantics, and information locality reflects collocational preferences.

Session 5B: Dialogue and Interactive Systems-10

"None of the Above": Measure Uncertainty in Dialog Response Retrieval [Website][PDF]
Yulan Feng, Shikib Mehri, Maxine Eskenazi, and Tiancheng Zhao 21:00–22:00

This paper discusses the importance of uncovering uncertainty in end-to-end dialog tasks and presents our experimental results on uncertainty classification on the processed Ubuntu Dialog Corpus. We show that instead of retraining models for this specific purpose, we can capture the original retrieval model's underlying confidence concerning the best prediction using trivial additional computation.

Can You Put it All Together: Evaluating Conversational Agents' Ability to Blend Skills [Website][PDF]
Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau 21:00–22:00

Being engaging, knowledgeable, and empathetic are all desirable general qualities in a conversational agent. Previous work has introduced tasks and datasets that aim to help agents to learn those qualities in isolation and gauge how well they can express them. But rather than being specialized in one single quality, a good open-domain conversational agent should be able to seamlessly blend them all into one cohesive conversational flow. In this work, we investigate several ways to combine models trained towards isolated capabilities, ranging from simple model aggregation schemes that require minimal additional training, to various forms of multi-task training that encompass several skills at all training stages. We further propose a new dataset, BlendedSkillTalk, to analyze how these capabilities would mesh together in a natural conversation, and compare the performance of different architectures and training schemes. Our experiments show that multi-tasking over several tasks that focus on particular capabilities results in better blended conversation performance compared to models trained on a single skill, and that both unified or two-stage approaches perform well if they are constructed to avoid unwanted bias in skill selection or are fine-tuned on our new task.

Conversational Graph Grounded Policy Learning for Open-Domain Conversation Generation [Website][PDF]
Jun Xu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu 21:00–22:00

To address the challenge of policy learning in open-domain multi-turn conversation, we propose to represent prior information about dialog transitions as a graph and learn a graph grounded dialog policy, aimed at fostering a more coherent and controllable dialog. To this end, we first construct a conversational graph (CG) from dialog corpora, in which there are vertices to represent "what to say" and "how to say", and edges to represent natural transition between a message (the last utterance in a dialog context) and its response. We then present a novel CG grounded policy learning framework that conducts dialog flow planning by graph traversal, which learns to identify a what-vertex and a how-vertex from the CG at each turn to guide response generation. In this way, we effectively leverage the CG to facilitate policy learning as follows: (1) it enables more effective long-term reward design, (2) it provides high-quality candidate actions, and (3) it gives us more control over the policy. Results on two benchmark corpora demonstrate the effectiveness of this framework.

Grounding Conversations with Improvised Dialogues [Website][PDF]
Hyundong Cho and Jonathan May 21:00–22:00

Effective dialogue involves grounding, the process of establishing mutual knowledge that is essential for communication between people. Modern dialogue systems are not explicitly trained to build common ground, and therefore overlook this important aspect of communication. Improvisational theater (improv) intrinsically contains a high proportion of dialogue focused on building common ground, and makes use of the yes-and principle, a strong grounding speech act, to establish coherence and an actionable objective reality. We collect a corpus of more than 26,000 yes-and turns, transcribing them from improv dialogues and extracting them from larger, but more sparsely populated movie script dialogue corpora, via a bootstrapped classifier. We fine-tune chit-chat dialogue systems with our corpus to encourage more grounded, relevant conversation and confirm these findings with human evaluations.

Image-Chat: Engaging Grounded Conversations [Website][PDF]
Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston 21:00–22:00

To achieve the long-term goal of machines being able to engage humans in conversation, our models should captivate the interest of their speaking partners. Communication grounded in images, whereby a dialogue is conducted based on a given photo, is a setup naturally appealing to humans (Hu et al., 2014). In this work we study large-scale architectures and datasets for this goal. We test a set of neural architectures using state-of-the-art image and text representations, considering various ways to fuse the components. To test such models, we collect a dataset of grounded human-human conversations, where speakers are asked to play roles given a provided emotional mood or style, as the use of such traits is also a key factor in engagingness (Guo et al., 2019). Our dataset, Image-Chat, consists of 202k dialogues over 202k images using 215 possible style traits. Automatic metrics and human evaluations of engagingness show the efficacy of our approach; in particular, we obtain state-of-the-art performance on the existing IGC task, and our best performing model is almost on par with humans on the Image-Chat test set (preferred 47.7% of the time).

Large Scale Multi-Actor Generative Dialog Modeling [Website][PDF]
Alex Boyd, Raul Puri, Mohammad Shoeybi, Mostofa Patwary, and Bryan Catanzaro 21:00–22:00

Non-goal oriented dialog agents (i.e. chatbots) aim to produce varying and engaging conversations with a user; however, they typically exhibit either inconsistent personality across conversations or the average personality of all users. This paper addresses these issues by controlling an agent's persona upon generation via conditioning on prior conversations of a target actor. In doing so, we are able to utilize more abstract patterns within a person's speech and better emulate them in generated responses. This work introduces the Generative Conversation Control model, an

augmented and fine-tuned GPT-2 language model that conditions on past reference conversations to probabilistically model multi-turn conversations in the actor's persona. We introduce an accompanying data collection procedure to obtain 10.3M conversations from 6 months worth of Reddit comments. We demonstrate that scaling model sizes from 117M to 8.3B parameters yields an improvement from 23.14 to 13.14 perplexity on 1.7M held out Reddit conversations. Increasing model scale yielded similar improvements in human evaluations that measure preference of model samples to the held out target distribution in terms of realism (31% increased to 37% preference), style matching (37% to 42%), grammar and content quality (29% to 42%), and conversation coherency (32% to 40%). We find that conditionally modeling past conversations improves perplexity by 0.47 in automatic evaluations. Through human trials we identify positive trends between conditional modeling and style matching and outline steps to further improve persona control.

Neural Generation of Dialogue Response Timings [Website][PDF]
Matthew Roddy and Naomi Harte 21:00–22:00

The timings of spoken response offsets in human dialogue have been shown to vary based on contextual elements of the dialogue. We propose neural models that simulate the distributions of these response offsets, taking into account the response turn as well as the preceding turn. The models are designed to be integrated into the pipeline of an incremental spoken dialogue system (SDS). We evaluate our models using offline experiments as well as human listening tests. We show that human listeners consider certain response timings to be more natural based on the dialogue context. The introduction of these models into SDS pipelines could increase the perceived naturalness of interactions.

[CL] The Design and Implementation of XiaoIce, an Empathetic Social Chatbot [Website][PDF]
Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum 21:00–22:00

This article describes the development of Microsoft XiaoIce, the most popular social chatbot in the world. XiaoIce is uniquely designed as an artifical intelligence companion with an emotional connection to satisfy the human need for communication, affection, and social belonging. We take into account both intelligent quotient and emotional quotient in system design, cast human–machine social chat as decision-making over Markov Decision Processes, and optimize XiaoIce for long-term user engagement, measured in expected Conversation-turns Per Session (CPS). We detail the system architecture and key components, including dialogue manager, core chat, skills, and an empathetic computing module. We show how XiaoIce dynamically recognizes human feelings and states, understands user intent, and responds to user needs throughout long conversations. Since the release in 2014, XiaoIce has communicated with over 660 million active users and succeeded in establishing long-term relationships with many of them. Analysis of large-scale online logs shows that XiaoIce has achieved an average CPS of 23, which is significantly higher than that of other chatbots and even human conversations.

The Dialogue Dodecathlon: Open-Domain Knowledge and Image Grounded Conversational Agents [Website][PDF]
Kurt Shuster, Da JU, Stephen Roller, Emily Dinan, Y-Lan Boureau, and Jason Weston 21:00–22:00

We introduce dodecaDialogue: a set of 12 tasks that measures if a conversational agent can communicate engagingly with personality and empathy, ask questions, answer questions by utilizing knowledge resources, discuss topics and situations, and perceive and converse about images. By multi-tasking on such a broad large-scale set of data, we hope to both move towards and measure progress in producing a single unified agent that can perceive, reason and converse with humans in an open-domain setting. We show that such multi-tasking improves over a BERT pre-trained baseline, largely due to multi-tasking with very large dialogue datasets in a similar domain, and that the multi-tasking in general provides gains to both text and image-based tasks using several metrics in both the fine-tune and task transfer settings. We obtain state-of-the-art results on many of the tasks, providing a strong baseline for this challenge.

Session 5B: Language Grounding to Vision, Robotics and Beyond-2

BabyWalk: Going Farther in Vision-and-Language Navigation by Taking Baby Steps [Website][PDF]
Wang Zhu, Hexiang Hu, Jiacheng Chen, Zhiwei Deng, Vihan Jain, Eugene Ie, and Fei Sha 21:00–22:00

Learning to follow instructions is of fundamental importance to autonomous agents for vision-and-language navigation (VLN). In this paper, we study how an agent can navigate long paths when learning from a corpus that consists of shorter ones. We show that existing state-of-the-art agents do not generalize well. To this end, we propose BabyWalk, a new VLN agent that is learned to navigate by decomposing long instructions into shorter ones (BabySteps) and completing them sequentially. A special design memory buffer is used by the agent to turn its past experiences into contexts for future steps. The learning process is composed of two phases. In the first phase, the agent uses imitation learning from demonstration to accomplish BabySteps. In the second phase, the agent uses curriculum-based reinforcement learning to maximize rewards on navigation tasks with increasingly longer instructions. We create two new benchmark datasets (of long navigation tasks) and use them in conjunction with existing ones to examine BabyWalk's generalization ability. Empirical results show that BabyWalk achieves state-of-the-art results on several metrics, in particular, is able to follow long instructions better. The codes and the datasets are released on our project page: https://github.com/Sha-Lab/babywalk.

Cross-media Structured Common Space for Multimedia Event Extraction [Website][PDF]
Manling Li, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, and Shih-Fu Chang 21:00–22:00

We introduce a new task, MultiMedia Event Extraction, which aims to extract events and their arguments from multimedia documents. We develop the first benchmark and collect a dataset of 245 multimedia news articles with extensively annotated events and arguments. We propose a novel method, Weakly Aligned Structured Embedding (WASE), that encodes structured representations of semantic information from textual and visual data into a common embedding space. The structures are aligned across modalities by employing a weakly supervised training strategy, which enables exploiting available resources without explicit cross-media annotation. Compared to uni-modal state-of-the-art methods, our approach achieves 4.0% and 9.8% absolute F-score gains on text event argument role labeling and visual event extraction. Compared to state-of-the-art multimedia unstructured representations, we achieve 8.3% and 5.0% absolute F-score gains on multimedia event extraction and argument role labeling, respectively. By utilizing images, we extract 21.4% more event mentions than traditional text-only methods.

Learning to Segment Actions from Observation and Narration [Website][PDF]
Daniel Fried, Jean-Baptiste Alayrac, Phil Blunsom, Chris Dyer, Stephen Clark, and Aida Nematzadeh 21:00–22:00

We apply a generative segmental model of task structure, guided by narration, to action segmentation in video. We focus on unsupervised and weakly-supervised settings where no action labels are known during training. Despite its simplicity, our model performs competitively with previous work on a dataset of naturalistic instructional videos. Our model allows us to vary the sources of supervision used in training, and we find that both task structure and narrative language provide large benefits in segmentation quality.

Learning to execute instructions in a Minecraft dialogue [Website][PDF]
Prashant Jayannavar, Anjali Narayan-Chen, and Julia Hockenmaier 21:00–22:00

The Minecraft Collaborative Building Task is a two-player game in which an Architect (A) instructs a Builder (B) to construct a target structure in a simulated Blocks World Environment. We define the subtask of predicting correct action sequences (block placements and removals) in a given game context, and show that capturing B's past actions as well as B's perspective leads to a significant improvement in performance on this challenging language understanding problem.

MART: Memory-Augmented Recurrent Transformer for Coherent Video Paragraph Captioning [Website][PDF]
Jie Lei, Liwei Wang, Yelong Shen, Dong Yu, Tamara Berg, and Mohit Bansal 21:00–22:00

Generating multi-sentence descriptions for videos is one of the most challenging captioning tasks due to its high requirements for not only visual relevance but also discourse-based coherence across the sentences in the paragraph. Towards this goal, we propose a new approach called Memory-Augmented Recurrent Transformer (MART), which uses a memory module to augment the transformer architecture. The memory module generates a highly summarized memory state from the video segments and the sentence history so as to help better prediction of the next sentence (w.r.t. coreference and repetition aspects), thus encouraging coherent paragraph generation. Extensive experiments, human evaluations, and qualitative analyses on two popular datasets ActivityNet Captions and YouCookII show that MART generates more coherent and less repetitive paragraph captions than baseline methods, while maintaining relevance to the input video events.

What is Learned in Visually Grounded Neural Syntax Acquisition [Website][PDF]
Noriyuki Kojima, Hadar Averbuch-Elor, Alexander Rush, and Yoav Artzi 21:00–22:00

Visual features are a promising signal for learning bootstrap textual models. However, blackbox learning models make it difficult to isolate the specific contribution of visual components. In this analysis, we consider the case study of the Visually Grounded Neural Syntax Learner (Shi et al., 2019), a recent approach for learning syntax from a visual training signal. By constructing simplified versions of the model, we isolate the core factors that yield the model's strong performance. Contrary to what the model might be capable of learning, we find significantly less expressive versions produce similar predictions and perform just as well, or even better. We also find that a simple lexical signal

of noun concreteness plays the main role in the model's predictions as opposed to more complex syntactic reasoning.

Session 5B: Machine Learning for NLP-4

A Batch Normalized Inference Network Keeps the KL Vanishing Away [Website][PDF]
Qile Zhu, Wei Bi, Xiaojiang Liu, Xiyao Ma, Xiaolin Li, and Dapeng Wu 21:00–22:00

Variational Autoencoder (VAE) is widely used as a generative model to approximate a model's posterior on latent variables by combining the amortized variational inference and deep neural networks. However, when paired with strong autoregressive decoders, VAE often converges to a degenerated local optimum known as "posterior collapse". Previous approaches consider the Kullback—Leibler divergence (KL) individual for each datapoint. We propose to let the KL follow a distribution across the whole dataset, and analyze that it is sufficient to prevent posterior collapse by keeping the expectation of the KL's distribution positive. Then we propose Batch Normalized-VAE (BN-VAE), a simple but effective approach to set a lower bound of the expectation by regularizing the distribution of the approximate posterior's parameters. Without introducing any new model component or modifying the objective, our approach can avoid the posterior collapse effectively and efficiently. We further show that the proposed BN-VAE can be extended to conditional VAE (CVAE). Empirically, our approach surpasses strong autoregressive baselines on language modeling, text classification and dialogue generation, and rivals more complex approaches while keeping almost the same training time as VAE.

Contextual Embeddings: When Are They Worth It? [Website][PDF]
Simran Arora, Avner May, Jian Zhang, and Christopher Ré 21:00–22:00

We study the settings for which deep contextual embeddings (e.g., BERT) give large improvements in performance relative to classic pretrained embeddings (e.g., GloVe), and an even simpler baseline—random word embeddings—focusing on the impact of the training set size and the linguistic properties of the task. Surprisingly, we find that both of these simpler baselines can match contextual embeddings on industry-scale data, and often perform within 5 to 10% accuracy (absolute) on benchmark tasks. Furthermore, we identify properties of data for which contextual embeddings give particularly large gains: language containing complex structure, ambiguous word usage, and words unseen in training.

[TACL] Efficient Contextual Representation Learning With Continuous Outputs [Website][PDF]
Liunian Harold Li, Patrick H. Chen, Cho-Jui Hsieh, and Kai-Wei Chang 21:00–22:00

Contextual representation models have achieved great success in improving various downstream natural language processing tasks. However, these language-model-based encoders are difficult to train due to their large parameter size and high computational complexity. By carefully examining the training procedure, we observe that the softmax layer, which predicts a distribution of the target word, often induces significant overhead, especially when the vocabulary size is large. Therefore, we revisit the design of the output layer and consider directly predicting the pre-trained embedding of the target word for a given context. When applied to ELMo, the proposed approach achieves a 4 times speedup and eliminates 80% trainable parameters while achieving competitive performance on downstream tasks. Further analysis shows that the approach maintains the speed advantage under various settings, even when the sentence encoder is scaled up.

Interactive Classification by Asking Informative Questions [Website][PDF]
Lili Yu, Howard Chen, Sida I. Wang, Tao Lei, and Yoav Artzi 21:00–22:00

We study the potential for interaction in natural language classification. We add a limited form of interaction for intent classification, where users provide an initial query using natural language, and the system asks for additional information using binary or multi-choice questions. At each turn, our system decides between asking the most informative question or making the final classification pre-diction. The simplicity of the model allows for bootstrapping of the system without interaction data, instead relying on simple crowd-sourcing tasks. We evaluate our approach on two domains, showing the benefit of interaction and the advantage of learning to balance between asking additional questions and making the final prediction.

Knowledge Graph Embedding Compression [Website][PDF]
Mrinmaya Sachan 21:00–22:00

Knowledge graph (KG) representation learning techniques that learn continuous embeddings of entities and relations in the KG have become popular in many AI applications. With a large KG, the embeddings consume a large amount of storage and memory. This is problematic and prohibits the deployment of these techniques in many real world settings. Thus, we propose an approach that compresses the KG embedding layer by representing each entity in the KG as a vector of discrete codes and then composes the embeddings from these codes. The approach can be trained end-to-end with simple modifications to any existing KG embedding technique. We evaluate the approach on various standard KG embedding evaluations and show that it achieves 50-1000x compression of embeddings with a minor loss in performance. The compressed embeddings also retain the ability to perform various reasoning tasks such as KG inference.

Low Resource Sequence Tagging using Sentence Reconstruction [Website][PDF]
Tal Perl, Sriram Chaudhury, and Raja Giryes 21:00–22:00

This work revisits the task of training sequence tagging models with limited resources using transfer learning. We investigate several proposed approaches introduced in recent works and suggest a new loss that relies on sentence reconstruction from normalized embeddings. Specifically, our method demonstrates how by adding a decoding layer for sentence reconstruction, we can improve the performance of various baselines. We show improved results on the CoNLL02 NER and UD 1.2 POS datasets and demonstrate the power of the method for transfer learning with low-resources achieving 0.6 F1 score in Dutch using only one sample from it.

Masked Language Model Scoring [Website][PDF]
Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff 21:00–22:00

Pretrained masked language models (MLMs) require finetuning for most NLP tasks. Instead, we evaluate MLMs out of the box via their pseudo-log-likelihood scores (PLLs), which are computed by masking tokens one by one. We show that PLLs outperform scores from autoregressive language models like GPT-2 in a variety of tasks. By rescoring ASR and NMT hypotheses, RoBERTa reduces an end-to-end LibriSpeech model's WER by 30% relative and adds up to +1.7 BLEU on state-of-the-art baselines for low-resource translation pairs, with further gains from domain adaptation. We attribute this success to PLL's unsupervised expression of linguistic acceptability without a left-to-right bias, greatly improving on scores from GPT-2 (+10 points on island effects, NPI licensing in BLiMP). One can finetune MLMs to give scores without masking, enabling computation in a single inference pass. In all, PLLs and their associated pseudo-perplexities (PPPLs) enable plug-and-play use of the growing number of pretrained MLMs; e.g., we use a single cross-lingual model to rescore translations in multiple languages. We release our library for language model scoring at https://github.com/awslabs/mlm-scoring.

Orthogonal Relation Transforms with Graph Context Modeling for Knowledge Graph Embedding [Website][PDF]
Yun Tang, Jing Huang, Guangtao Wang, Xiaodong He, and Bowen Zhou 21:00–22:00

Distance-based knowledge graph embeddings have shown substantial improvement on the knowledge graph link prediction task, from TransE to the latest state-of-the-art RotatE. However, complex relations such as N-to-1, 1-to-N and N-to-N still remain challenging to predict. In this work, we propose a novel distance-based approach for knowledge graph link prediction. First, we extend the RotatE from 2D complex domain to high dimensional space with orthogonal transforms to model relations. The orthogonal transform embedding for relations keeps the capability for modeling symmetric/anti-symmetric, inverse and compositional relations while achieves better modeling capacity. Second, the graph context is integrated into distance scoring functions directly. Specifically, graph context is explicitly modeled via two directed context representations. Each node embedding in knowledge graph is augmented with two context representations, which are computed from the neighboring outgoing and incoming nodes/edges respectively. The proposed approach improves prediction accuracy on the difficult N-to-1, 1-to-N and N-to-N cases. Our experimental results show that it achieves state-of-the-art results on two common benchmarks FB15k-237 and WNRR-18, especially on FB15k-237 which has many high in-degree nodes.

[TACL] Perturbation Based Learning for Structured NLP Tasks with Application to Dependency Parsing [Website][PDF]
Amichay Doitch, Ram Yazdi, Tamir Hazan, and Roi Reichart 21:00–22:00

The best solution of structured prediction models in NLP is often inaccurate due to limited expressive power of the model or to non-exact parameter estimation. One way to mitigate this problem is sampling candidate solutions from the model's solution space, reasoning that effective exploration of this space should yield high quality solutions. Unfortunately, sampling is often computationally hard and many works hence back-off to sub-optimal strategies such as extraction of the best scoring solutions of the model, which are not as diverse as sampled solutions. In this paper we propose a perturbation-based approach where sampling from a probabilistic model is computationally efficient. We present a learning algorithm for the variance of the perturbations, and empirically demonstrate its importance. Moreover, while finding the argmax in our model is intractable, we propose an efficient and effective approximation. We apply our framework to cross-lingual dependency parsing across 72 corpora from 42 languages and to lightly supervised dependency parsing across 13 corpora from 12 languages and demonstrate strong results in terms of both the quality of the entire solution list and of the final solution.

Posterior Calibrated Training on Sentence Classification Tasks [Website][PDF]
Taehee Jung, Dongyeop Kang, Hua Cheng, Lucas Mentch, and Thomas Schaaf 21:00–22:00

Most classification models work by first predicting a posterior probability distribution over all classes and then selecting that class with the largest estimated probability. In many settings however, the quality of posterior probability itself (e.g., 65% chance having diabetes), gives more reliable information than the final predicted class alone. When these methods are shown to be poorly calibrated, most fixes to date have relied on posterior calibration, which rescales the predicted probabilities but often has little impact on final classifications. Here we propose an end-to-end training procedure called posterior calibrated (PosCal) training that directly optimizes the objective while minimizing the difference between the predicted and empirical posterior probabilities.We show that PosCal not only helps reduce the calibration error but also improve task performance by penalizing drops in performance of both objectives. Our PosCal achieves about 2.5% of task performance gain and 16.1% of calibration error reduction on GLUE (Wang et al., 2018) compared to the baseline. We achieved the comparable task performance with 13.2% calibration error reduction on xSLUE (Kang and Hovy, 2019), but not outperforming the two-stage calibration baseline. PosCal training can be easily extendable to any types of classification tasks as a form of regularization term. Also, PosCal has the advantage that it incrementally tracks needed statistics for the calibration objective during the training process, making efficient use of large training sets.

Posterior Control of Blackbox Generation [Website][PDF]
Xiang Lisa Li and Alexander Rush 21:00–22:00

Text generation often requires high-precision output that obeys task-specific rules. This fine-grained control is difficult to enforce with off-the-shelf deep learning models. In this work, we consider augmenting neural generation models with discrete control states learned through a structured latent-variable approach. Under this formulation, task-specific knowledge can be encoded through a range of rich, posterior constraints that are effectively trained into the model. This approach allows users to ground internal model decisions based on prior knowledge, without sacrificing the representational power of neural generative models. Experiments consider applications of this approach

for text generation. We find that this method improves over standard benchmarks, while also providing fine-grained control.

Pretrained Transformers Improve Out-of-Distribution Robustness [Website][PDF]
Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song 21:00–22:00

Although pretrained Transformers such as BERT achieve high accuracy on in-distribution examples, do they generalize to new distributions? We systematically measure out-of-distribution (OOD) generalization for seven NLP datasets by constructing a new robustness benchmark with realistic distribution shifts. We measure the generalization of previous models including bag-of-words models, ConvNets, and LSTMs, and we show that pretrained Transformers' performance declines are substantially smaller. Pretrained transformers are also more effective at detecting anomalous or OOD examples, while many previous models are frequently worse than chance. We examine which factors affect robustness, finding that larger models are not necessarily more robust, distillation can be harmful, and more diverse pretraining data can enhance robustness. Finally, we show where future work can improve OOD robustness.

Robust Encodings: A Framework for Combating Adversarial Typos [Website][PDF]
Erik Jones, Robin Jia, Aditi Raghunathan, and Percy Liang 21:00–22:00

Despite excellent performance on many tasks, NLP systems are easily fooled by small adversarial perturbations of inputs. Existing procedures to defend against such perturbations are either (i) heuristic in nature and susceptible to stronger attacks or (ii) provide guaranteed robustness to worst-case attacks, but are incompatible with state-of-the-art models like BERT. In this work, we introduce robust encodings (RobEn): a simple framework that confers guaranteed robustness, without making compromises on model architecture. The core component of RobEn is an encoding function, which maps sentences to a smaller, discrete space of encodings. Systems using these encodings as a bottleneck confer guaranteed robustness with standard training, and the same encodings can be used across multiple tasks. We identify two desiderata to construct robust encoding functions: perturbations of a sentence should map to a small set of encodings (stability), and models using encodings should still perform well (fidelity). We instantiate RobEn to defend against a large family of adversarial typos. Across six tasks from GLUE, our instantiation of RobEn paired with BERT achieves an average robust accuracy of 71.3% against all adversarial typos in the family considered, while previous work using a typo-corrector achieves only 35.3% accuracy against a simple greedy attack.

Showing Your Work Doesn't Always Work [Website][PDF]
Raphael Tang, Jaejun Lee, Ji Xin, Xinyu Liu, Yaoliang Yu, and Jimmy Lin 21:00–22:00

In natural language processing, a recently popular line of work explores how to best report the experimental results of neural networks. One exemplar publication, titled "Show Your Work: Improved Reporting of Experimental Results" (Dodge et al., 2019), advocates for reporting the expected validation effectiveness of the best-tuned model, with respect to the computational budget. In the present work, we critically examine this paper. As far as statistical generalizability is concerned, we find unspoken pitfalls and caveats with this approach. We analytically show that their estimator is biased and uses error-prone assumptions. We find that the estimator favors negative errors and yields poor bootstrapped confidence intervals. We derive an unbiased alternative and bolster our claims with empirical evidence from statistical simulation. Our codebase is at https://github.com/castorini/meanmax.

Span Selection Pre-training for Question Answering [Website][PDF]
Michael Glass, Alfio Gliozzo, Rishav Chakravarti, Anthony Ferritto, Lin Pan, G P Shrivatsa Bhargav, Dinesh Garg, and Avi Sil 21:00–22:00

BERT (Bidirectional Encoder Representations from Transformers) and related pre-trained Transformers have provided large gains across many language understanding tasks, achieving a new state-of-the-art (SOTA). BERT is pre-trained on two auxiliary tasks: Masked Language Model and Next Sentence Prediction. In this paper we introduce a new pre-training task inspired by reading comprehension to better align the pre-training from memorization to understanding. Span Selection PreTraining (SSPT) poses cloze-like training instances, but rather than draw the answer from the model's parameters, it is selected from a relevant passage. We find significant and consistent improvements over both BERT-BASE and BERT-LARGE on multiple Machine Reading Comprehension (MRC) datasets. Specifically, our proposed model has strong empirical evidence as it obtains SOTA results on Natural Questions, a new benchmark MRC dataset, outperforming BERT-LARGE by 3 F1 points on short answer prediction. We also show significant impact in HotpotQA, improving answer prediction F1 by 4 points and supporting fact prediction F1 by 1 point and outperforming the previous best system. Moreover, we show that our pre-training approach is particularly effective when training data is limited, improving the learning curve by a large amount.

Topological Sort for Sentence Ordering [Website][PDF]
Shrimai Prabhumoye, Ruslan Salakhutdinov, and Alan W Black 21:00–22:00

Sentence ordering is the task of arranging the sentences of a given text in the correct order. Recent work using deep neural networks for this task has framed it as a sequence prediction problem. In this paper, we propose a new framing of this task as a constraint solving problem and introduce a new technique to solve it. Additionally, we propose a human evaluation for this task. The results on both automatic and human metrics across four different datasets show that this new technique is better at capturing coherence in documents.

Weight Poisoning Attacks on Pretrained Models [Website][PDF]
Keita Kurita, Paul Michel, and Graham Neubig 21:00–22:00

Recently, NLP has seen a surge in the usage of large pre-trained models. Users download weights of models pre-trained on large datasets, then fine-tune the weights on a task of their choice. This raises the question of whether downloading untrusted pre-trained weights can pose a security threat. In this paper, we show that it is possible to

construct "weight poisoning" attacks where pre-trained weights are injected with vulnerabilities that expose "backdoors" after fine-tuning, enabling the attacker to manipulate the model prediction simply by injecting an arbitrary keyword. We show that by applying a regularization method which we call RIPPLe and an initialization procedure we call Embedding Surgery, such attacks are possible even with limited knowledge of the dataset and fine-tuning procedure. Our experiments on sentiment classification, toxicity detection, and spam detection show that this attack is widely applicable and poses a serious threat. Finally, we outline practical defenses against such attacks.[2]

schuBERT: Optimizing Elements of BERT [Website] [PDF]
Ashish Khetan and Zohar Karnin 21:00–22:00

Transformers have gradually become a key component for many state-of-the-art natural language representation models. A recent Transformer based model- BERTachieved state-of-the-art results on various natural language processing tasks, including GLUE, SQuAD v1.1, and SQuAD v2.0. This model however is computationally prohibitive and has a huge number of parameters. In this work we revisit the architecture choices of BERT in efforts to obtain a lighter model. We focus on reducing the number of parameters yet our methods can be applied towards other objectives such FLOPs or latency. We show that much efficient light BERT models can be obtained by reducing algorithmically chosen correct architecture design dimensions rather than reducing the number of Transformer encoder layers. In particular, our schuBERT gives 6.6% higher average accuracy on GLUE and SQuAD datasets as compared to BERT with three encoder layers while having the same number of parameters.

[2] Our code will be made publicly available on publication.

Session 5B: NLP Applications-4

"The Boating Store Had Its Best Sail Ever": Pronunciation-attentive Contextualized Pun Recognition [Website][PDF]
Yichao Zhou, Jyun-Yu Jiang, Jieyu Zhao, Kai-Wei Chang, and Wei Wang 21:00–22:00

Humor plays an important role in human languages and it is essential to model humor when building intelligence systems. Among different forms of humor, puns perform wordplay for humorous effects by employing words with double entendre and high phonetic similarity. However, identifying and modeling puns are challenging as puns usually involved implicit semantic or phonological tricks. In this paper, we propose Pronunciation-attentive Contextualized Pun Recognition (PCPR) to perceive human humor, detect if a sentence contains puns and locate them in the sentence. PCPR derives contextualized representation for each word in a sentence by capturing the association between the surrounding context and its corresponding phonetic symbols. Extensive experiments are conducted on two benchmark datasets. Results demonstrate that the proposed approach significantly outperforms the state-of-the-art methods in pun detection and location tasks. In-depth analyses verify the effectiveness and robustness of PCPR.

A Girl Has A Name: Detecting Authorship Obfuscation [Website][PDF]
Asad Mahmood, Zubair Shafiq, and Padmini Srinivasan 21:00–22:00

Authorship attribution aims to identify the author of a text based on the stylometric analysis. Authorship obfuscation, on the other hand, aims to protect against authorship attribution by modifying a text's style. In this paper, we evaluate the stealthiness of state-of-the-art authorship obfuscation methods under an adversarial threat model. An obfuscator is stealthy to the extent an adversary finds it challenging to detect whether or not a text modified by the obfuscator is obfuscated — a decision that is key to the adversary interested in authorship attribution. We show that the existing authorship obfuscation methods are not stealthy as their obfuscated texts can be identified with an average F1 score of 0.87. The reason for the lack of stealthiness is that these obfuscators degrade text smoothness, as ascertained by neural language models, in a detectable manner. Our results highlight the need to develop stealthy authorship obfuscation methods that can better protect the identity of an author seeking anonymity.

DeeBERT: Dynamic Early Exiting for Accelerating BERT Inference [Website][PDF]
Ji Xin, Raphael Tang, Jaejun Lee, Yaoliang Yu, and Jimmy Lin 21:00–22:00

Large-scale pre-trained language models such as BERT have brought significant improvements to NLP applications. However, they are also notorious for being slow in inference, which makes them difficult to deploy in real-time applications. We propose a simple but effective method, DeeBERT, to accelerate BERT inference. Our approach allows samples to exit earlier without passing through the entire model. Experiments show that DeeBERT is able to save up to ~40% inference time with minimal degradation in model quality. Further analyses show different behaviors in the BERT transformer layers and also reveal their redundancy. Our work provides new ideas to efficiently apply deep transformer-based models to downstream tasks. Code is available at https://github.com/castorini/DeeBERT.

Efficient Strategies for Hierarchical Text Classification: External Knowledge and Auxiliary Tasks [Website][PDF]
Kervy Rivas Rojas, Gina Bustamante, Arturo Oncevay, and Marco Antonio Sobrevilla Cabezudo 21:00–22:00

In hierarchical text classification, we perform a sequence of inference steps to predict the category of a document from top to bottom of a given class taxonomy. Most of the studies have focused on developing novels neural network architectures to deal with the hierarchical structure, but we prefer to look for efficient ways to strengthen a baseline model. We first define the task as a sequence-to-sequence problem. Afterwards, we propose an auxiliary synthetic task of bottom-up-classification. Then, from external dictionaries, we retrieve textual definitions for the classes of all the hierarchy's layers, and map them into the word vector space. We use the class-definition embeddings as an additional input to condition the prediction of the next layer and in an adapted beam search. Whereas the modified search did not provide large gains, the combination of the auxiliary task and the additional input of class-definitions significantly enhance the classification accuracy. With our efficient approaches, we outperform previous studies, using a drastically reduced number of parameters, in two well-known English datasets.

Investigating the effect of auxiliary objectives for the automated grading of learner English speech transcriptions [Website][PDF]
Hannah Craighead, Andrew Caines, Paula Buttery, and Helen Yannakoudakis 21:00–22:00

We address the task of automatically grading the language proficiency of spontaneous speech based on textual features from automatic speech recognition transcripts. Motivated by recent advances in multi-task learning, we develop neural networks trained in a multi-task fashion that learn to predict the proficiency level of non-native English speakers by taking advantage of inductive transfer between the main task (grading) and auxiliary prediction tasks: morpho-syntactic labeling, language modeling, and native language identification (L1). We encode the transcriptions with both bi-directional recurrent neural networks and with bi-directional representations from transformers, compare against a feature-rich baseline, and analyse performance at different proficiency levels and with transcriptions of varying error rates. Our best performance comes from a transformer encoder with L1 prediction as an auxiliary task. We discuss areas for improvement and potential applications for text-only speech scoring.

SPECTER: Document-level Representation Learning using Citation-informed Transformers [Website][PDF]
Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld 21:00–22:00

Representation learning is a critical ingredient for natural language processing systems. Recent Transformer language models like BERT learn powerful textual representations, but these models are targeted towards token- and sentence-level training objectives and do not leverage information on inter-document relatedness, which limits their document-level representation power. For applications on scientific documents, such as classification and recommendation, accurate embeddings of documents are a necessity. We propose SPECTER, a new method to generate document-level embedding of scientific papers based on pretraining a Transformer language model on a powerful signal of document-level relatedness: the citation graph. Unlike existing pretrained language models, Specter can be easily applied to downstream applications without task-specific fine-tuning. Additionally, to encourage further research on document-level models, we introduce SciDocs, a new evaluation benchmark consisting of seven document-level tasks ranging from citation prediction, to document classification and recommendation. We show that Specter outperforms a variety of competitive baselines on the benchmark.

Semantic Scaffolds for Pseudocode-to-Code Generation [Website] [PDF]
Ruiqi Zhong, Mitchell Stern, and Dan Klein 21:00–22:00

We propose a method for program generation based on semantic scaffolds, lightweight structures representing the high-level semantic and syntactic composition of a program. By first searching over plausible scaffolds then using these as constraints for a beam search over programs, we achieve better coverage of the search space when compared with existing techniques. We apply our hierarchical search method to the SPoC dataset for pseudocode-to-code generation, in which we are given line-level natural language pseudocode annotations and aim to produce a program satisfying execution-based test cases. By using semantic scaffolds during inference, we achieve a 10% absolute improvement in top-100 accuracy over the previous state-of-the-art. Additionally, we require only 11 candidates to reach the top-3000 performance of the previous best approach when tested against unseen problems, demonstrating a substantial improvement in efficiency.

Session 5B Semantics: Lexical-3

Breaking Through the 80% Glass Ceiling: Raising the State of the Art in Word Sense Disambiguation by Incorporating Knowledge Graph Information [Website][PDF]
Michele Bevilacqua and Roberto Navigli 21:00–22:00

Neural architectures are the current state of the art in Word Sense Disambiguation (WSD). However, they make limited use of the vast amount of relational information encoded in Lexical Knowledge Bases (LKB). We present Enhanced WSD Integrating Synset Embeddings and Relations (EWISER), a neural supervised architecture that is able to tap into this wealth of knowledge by embedding information from the LKB graph within the neural architecture, and to exploit pretrained synset embeddings, enabling the network to predict synsets that are not in the training set. As a result, we set a new state of the art on almost all the evaluation settings considered, also breaking through, for the first time, the 80% ceiling on the concatenation of all the standard all-words English WSD evaluation benchmarks. On multilingual all-words WSD, we report state-of-the-art results by training on nothing but English.

Glyph2Vec: Learning Chinese Out-of-Vocabulary Word Embedding from Glyphs [Website][PDF]
Hong-You Chen, SZ-HAN YU, and Shou-de Lin 21:00–22:00

Chinese NLP applications that rely on large text often contain huge amounts of vocabulary which are sparse in corpus. We show that characters' written form, *Glyphs*, in ideographic languages could carry rich semantics. We present a multi-modal model, *Glyph2Vec*, to tackle Chinese out-of-vocabulary word embedding problem. Glyph2Vec extracts visual features from word glyphs to expand current word embedding space for out-of-vocabulary word embedding, without the need of accessing any corpus, which is useful for improving Chinese NLP systems, especially for low-resource scenarios. Experiments across different applications show the significant effectiveness of our model.

[TACL] Learning Lexical Subspaces in a Distributional Vector Space [Website][PDF]
Kushal Arora, Aishik Chakraborty, and Jackie Chi Kit Cheung 21:00–22:00

In this paper, we propose LexSub, a novel approach towards unifying lexical and distributional semantics. We inject knowledge about lexical-semantic relations into distributional word embeddings by defining subspaces of the distributional vector space in which a lexical relation should hold. Our framework can handle symmetric attract and repel relations (e.g., synonymy and antonymy, respectively), as well as asymmetric relations (e.g., hypernymy and meronomy). In a suite of intrinsic benchmarks, we show that our model outperforms previous approaches on relatedness tasks and on hypernymy classification and detection, while being competitive on word similarity tasks. It also outperforms previous systems on extrinsic classification tasks that benefit from exploiting lexical relational cues. We perform a series of analyses to understand the behaviors of our model.

Multidirectional Associative Optimization of Function-Specific Word Representations [Website][PDF]
Daniela Gerz, Ivan Vulić, Marek Rei, Roi Reichart, and Anna Korhonen 21:00–22:00

We present a neural framework for learning associations between interrelated groups of words such as the ones found in Subject-Verb-Object (SVO) structures. Our model induces a joint function-specific word vector space, where vectors of e.g. plausible SVO compositions lie close together. The model retains information about word group membership even in the joint space, and can thereby effectively be applied to a number of tasks reasoning over the SVO structure. We show the robustness and versatility of the proposed framework by reporting state-of-the-art results on the tasks of estimating selectional preference and event similarity. The results indicate that the combinations of representations learned with our task-independent model outperform task-specific architectures from prior work, while reducing the number of parameters by up to 95%.

Predicting Degrees of Technicality in Automatic Terminology Extraction [Website][PDF]
Anna Hätty, Dominik Schlechtweg, Michael Dorna, and Sabine Schulte im Walde 21:00–22:00

While automatic term extraction is a well-researched area, computational approaches to distinguish between degrees of technicality are still understudied. We semi-automatically create a German gold standard of technicality across four domains, and illustrate the impact of a web-crawled general-language corpus on technicality prediction. When defining a classification approach that combines general-language and domain-specific word embeddings, we go beyond previous work and align vector spaces to gain comparative embeddings. We suggest two novel models to exploit general- vs. domain-specific comparisons: a simple neural network model with pre-computed comparative-embedding information as input, and a multi-channel model computing the comparison internally. Both models outperform previous approaches, with the multi-channel model performing best.

Verbal Multiword Expressions for Identification of Metaphor [Website][PDF]
Omid Rohanian, Marek Rei, Shiva Taslimipoor, and Le An Ha 21:00–22:00

Metaphor is a linguistic device in which a concept is expressed by mentioning another. Identifying metaphorical expressions, therefore, requires a non-compositional understanding of semantics. Multiword Expressions (MWEs), on the other hand, are linguistic phenomena with varying degrees of semantic opacity and their identification poses a challenge to computational models. This work is the first attempt at analysing the interplay of metaphor and MWEs processing through the design of a neural architecture whereby classification of metaphors is enhanced by informing the model of the presence of MWEs. To the best of our knowledge, this is the first "MWE-aware" metaphor identification system paving the way for further experiments on the complex interactions of these phenomena. The results and analyses show that this proposed architecture reach state-of-the-art on two different established metaphor datasets.

Demo Session 5C

Time: *21:30–22:15*

jiant: A Software Toolkit for Research on General-Purpose Text Understanding Models [Website][PDF]
Yada Pruksachatkun, Phil Yeres, Haokun Liu, Jason Phang, Phu Mon Htut, Alex Wang, Ian Tenney, and Samuel R. Bowman

We introduce jiant, an open source toolkit for conducting multitask and transfer learning experiments on English NLU tasks. jiant enables modular and configuration driven experimentation with state-of-the-art models and a broad set of tasks for probing, transfer learning, and multitask training experiments. jiant implements over 50 NLU tasks, including all GLUE and SuperGLUE benchmark tasks. We demonstrate that jiant reproduces published performance on a variety of tasks and models, e.g., RoBERTa and BERT.

The Microsoft Toolkit of Multi-Task Deep Neural Networks for Natural Language Understanding [Website][PDF]
Xiaodong Liu, Yu Wang, Jianshu Ji, Hao Cheng, Xueyun Zhu, Emmanuel Awa, Pengcheng He, Weizhu Chen, Hoifung Poon, Guihong Cao, and Jianfeng Gao

We present MT-DNN, an open-source natural language understanding (NLU) toolkit that makes it easy for researchers and developers to train customized deep learning models. Built upon PyTorch and Transformers, MT-DNN is designed to facilitate rapid customization for a broad spectrum of NLU tasks, using a variety of objectives (classification, regression, structured prediction) and text encoders (e.g., RNNs, BERT, RoBERTa, UniLM). A unique feature of MT-DNN is its built-in support for robust and transferable learning using the adversarial multi-task learning paradigm. To enable efficient production deployment, MT-DNN supports multi-task knowledge distillation, which can substantially compress a deep neural model without significant performance drop. We demonstrate the effectiveness of MT-DNN on a wide range of NLU applications across general and biomedical domains. The software and pre-trained models will be publicly available at https://github.com/namisan/mt-dnn.

Stanza: A Python Natural Language Processing Toolkit for Many Human Languages [Website][PDF]
Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning

We introduce Stanza, an open-source Python natural language processing toolkit supporting 66 human languages. Compared to existing widely used toolkits, Stanza features a language-agnostic fully neural pipeline for text analysis, including tokenization, multi-word token expansion, lemmatization, part-of-speech and morphological feature tagging, dependency parsing, and named entity recognition. We have trained Stanza on a total of 112 datasets, including the Universal Dependencies treebanks and other multilingual corpora, and show that the same neural architecture generalizes well and achieves competitive performance on all languages tested. Additionally, Stanza includes a native Python interface to the widely used Java Stanford CoreNLP software, which further extends its functionality to cover other tasks such as coreference resolution and relation extraction. Source code, documentation, and pretrained models for 66 languages are available at https://stanfordnlp.github.io/stanza/.

4

Main Conference: Tuesday, July 7

Overview

5:00–5:45 **Demo Session 1A**

5:00–6:00 **Session 6A**
Ethics and NLP-1
Machine Learning for NLP-5
Machine Translation-7
NLP Applications-5
Sentiment Analysis, Stylistic Analysis, and Argument Mining-1
Student Research Workshop
Tagging, Chunking and Parsing-1

5:45–6:30 **Demo Session 1B**

6:00–7:00 **Session 6B**
Computational Social Science and Social Media-4
Interpretability and Analysis of Models for NLP-1
Machine Learning for NLP-6
Machine Translation-8
Resources and Evaluation-5
Lexical-4
Sentiment Analysis, Stylistic Analysis, and Argument Mining-2
Speech and Multimodality-3
Student Research Workshop

6:30–7:15 **Demo Session 1C**

8:00–8:45 **Demo Session 2A**

8:00–9:00 **Session 7A**
Computational Social Science and Social Media-5
Generation-9
Machine Learning for NLP-7
Machine Translation-9
Question Answering-4
Resources and Evaluation-6
Lexical-5
Sentiment Analysis, Stylistic Analysis, and Argument Mining-3
Student Research Workshop
Tagging, Chunking and Parsing-2

8:45–9:30 **Demo Session 2B**

9:00–10:00 **Session 7B**
Ethics and NLP-2
Interpretability and Analysis of Models for NLP-2
Machine Learning for NLP-8
Machine Translation-10
NLP Applications-6
Sentence Level-3
Sentiment Analysis, Stylistic Analysis, and Argument Mining-4
Speech and Multimodality-4
Student Research Workshop

12:00–12:45 **Demo Session 3A**

12:00–13:00 **Session 8A**
Computational Social Science and Social Media-6
Interpretability and Analysis of Models for NLP-3
Machine Translation-11
Question Answering-5
Resources and Evaluation-7
Lexical-6
Sentence Level-4
Sentiment Analysis, Stylistic Analysis, and Argument Mining-5
Student Research Workshop
Tagging, Chunking and Parsing-3

12:45–13:30 **Demo Session 3B**

13:00–14:00 **Session 8B**
Ethics and NLP-3
Generation-10
Interpretability and Analysis of Models for NLP-4
Machine Learning for NLP-9
Machine Translation-12
NLP Applications-7
Resources and Evaluation-8
Sentiment Analysis, Stylistic Analysis, and Argument Mining-6
Speech and Multimodality-5
Student Research Workshop

13:30–14:15 **Demo Session 3C**

14:00–14:30 **Lifetime Achievement Award Video Livestream (Sponsored by Bloomberg Engineering and IBM Research AI)**

14:30–14:45 **Lifetime Achievement Award Live Q&A (Sponsored by Bloomberg Engineering and IBM Research AI)**

14:45–15:15 **Distinguished Service Award, Test-of-Time Award Video and Q&A**

15:15–16:00 **Reviewing Meeting Q&A**

17:00–17:45 **Demo Session 4A**

17:00–18:00 **Session 9A**
Dialogue and Interactive Systems-11
Interpretability and Analysis of Models for NLP-5
Language Grounding to Vision, Robotics and Beyond-3
Machine Learning for NLP-10
Resources and Evaluation-9
Student Research Workshop
Summarization-4
Theme-1

17:45–18:30 **Demo Session 4B**

18:00–19:00 **Session 9B**
Computational Social Science and Social Media-7
Discourse and Pragmatics-4
Ethics and NLP-4
Interpretability and Analysis of Models for NLP-6
Question Answering-6
Resources and Evaluation-10
Sentiment Analysis, Stylistic Analysis, and Argument Mining-7
Student Research Workshop

18:30–19:15 **Demo Session 4C**

20:00–20:45 **Demo Session 5A**

20:00–21:00 **Session 10A**
Computational Social Science and Social Media-8
Dialogue and Interactive Systems-12
Interpretability and Analysis of Models for NLP-7
Question Answering-7
Resources and Evaluation-11
Sentiment Analysis, Stylistic Analysis, and Argument Mining-8
Theme-2

20:45–21:30 **Demo Session 5B**

21:00–22:00 **Session 10B**
Discourse and Pragmatics-5
Ethics and NLP-5
Interpretability and Analysis of Models for NLP-8
Language Grounding to Vision, Robotics and Beyond-4
Machine Learning for NLP-11
Question Answering-8
Resources and Evaluation-12
Speech and Multimodality-6
Summarization-5

21:30–22:15 **Demo Session 5C**

Demo Session 1A

Time: *5:00–5:45*

LinggleWrite: a Coaching System for Essay Writing [Website][PDF]
Chung-Ting Tsai, Jhih-Jie Chen, Ching-Yu Yang, and Jason S. Chang

This paper presents LinggleWrite, a writing coach that provides writing suggestions, assesses writing proficiency levels, detects grammatical errors, and offers corrective feedback in response to user's essay. The method involves extracting grammar patterns, training models for automated essay scoring (AES) and grammatical error detection (GED), and finally retrieving plausible corrections from a n-gram search engine. Experiments on public test sets indicate that both AES and GED models achieve state-of-the-art performance. These results show that LinggleWrite is potentially useful in helping learners improve their writing skills.

Session 6A Overview – Tuesday, July 7, 2020 5:00–6:00

Track A *Ethics and NLP-1* Abstracts	Gender Bias in Multilingual Embeddings and Cross-Lingual Transfer *Zhao, Mukherjee, Hosseini, Chang, and Hassan Awadallah* [Website] [PDF]	Give Me Convenience and Give Her Death: Who Should Decide What Uses of NLP are Appropriate, and on What Basis? *, Lau, and Baldwin* [Website] [PDF]	Is Your Classifier Actually Biased? Measuring Fairness under Uncertainty with Bernstein Bounds *Ethayarajh* [Website] [PDF]	It's Morphin' Time! Combating Linguistic Discrimination with Inflectional Perturbations *Tan, Joty, Kan, and Socher* [Website] [PDF]	Mitigating Gender Bias Amplification in Distribution by Posterior Regularization *Jia, Meng, Zhao, and Chang* [Website] [PDF]
	Towards Understanding Gender Bias in Relation Extraction *Gaut, Sun, Tang, Huang, Qian, ElSherief, Zhao, Mirza, Belding, Chang, and Wang* [Website] [PDF]				
Track B *Machine Learning for NLP-5* Abstracts	A Probabilistic Generative Model for Typographical Analysis of Early Modern Printing *Goyal, Dyer, Warren, G'Sell, and Berg-Kirkpatrick* [Website] [PDF]	Attentive Pooling with Learnable Norms for Text Representation *Wu, Wu, Qi, Cui, and Huang* [Website] [PDF]	Estimating the influence of auxiliary tasks for multi-task learning of sequence tagging tasks *Schröder and Biemann* [Website] [PDF]	How Does Selective Mechanism Improve Self-Attention Networks? *Geng, Wang, Wang, Qin, Liu, and Tu* [Website] [PDF]	Improving Transformer Models by Reordering their Sublayers *Press, Smith, and Levy* [Website] [PDF]
	Single Model Ensemble using Pseudo-Tags and Distinct Vectors *Kuwabara, Suzuki, and Nakayama* [Website] [PDF]	Zero-shot Text Classification via Reinforced Self-training *Ye, Geng, Chen, Chen, Xu, Zheng, Wang, Zhang, and Chen* [Website] [PDF]			
Track C *Machine Translation-7* Abstracts	A Novel Graph-based Multi-modal Fusion Encoder for Neural Machine Translation *Yin, Meng, Su, Zhou, Yang, Zhou, and Luo* [Website] [PDF]	A Relaxed Matching Procedure for Unsupervised BLI *Zhao, Wang, Zhang, and Wu* [Website] [PDF]	Dynamic Programming Encoding for Subword Segmentation in Neural Machine Translation *He, Haffari, and Norouzi* [Website] [PDF]	Geometry-aware domain adaptation for unsupervised alignment of word embeddings *Jawanpuria, Meghwanshi, and Mishra* [Website] [PDF]	Learning to Recover from Multi-Modality Errors for Non-Autoregressive Neural Machine Translation *Ran, Lin, Li, and Zhou* [Website] [PDF]
	On the Inference Calibration of Neural Machine Translation *Wang, Tu, Shi, and Liu* [Website] [PDF]	[CL] Unsupervised Word Translation with Adversarial Autoencoder *Mohiuddin and Joty* [Website] [PDF]			
Track D *NLP Applications-5* Abstracts	Camouflaged Chinese Spam Content Detection with Semi-supervised Generative Active Learning *Jiang, Gao, Duan, Kang, Sun, Zhang, and Liu* [Website] [PDF]	Distinguish Confusing Law Articles for Legal Judgment Prediction *Xu, Wang, Chen, Pan, Wang, and Zhao* [Website] [PDF]	Hiring Now: A Skill-Aware Multi-Attention Model for Job Posting Generation *Liu, Liu, Zhang, Chi, Shi, and Huang* [Website] [PDF]	HyperCore: Hyperbolic and Co-graph Representation for Automatic ICD Coding *Cao, Chen, Liu, Zhao, Liu, and Chong* [Website] [PDF]	Hyperbolic Capsule Networks for Multi-Label Classification *Chen, Huang, Xiao, and Jing* [Website] [PDF]

	Improving Segmentation for Technical Support Problems *Chauhan and Gupta* [Website][PDF]	MOOCCube: A Large-scale Data Repository for NLP Applications in MOOCs *Yu, Luo, Xiao, Zhong, Wang, Luo, Wang, Hou, Li, Liu, and Tang* [Website][PDF]	Towards Interpretable Clinical Diagnosis with Bayesian Network Ensembles Stacked on Entity-Aware CNNs *Chen, Dai, Yuan, Lu, and Huang* [Website][PDF]		
Track E *Sentiment Analysis, Stylistic Analysis, and Argument Mining-1* Abstracts	Analyzing the Persuasive Effect of Style in News Editorial Argumentation *El Baff, Wachsmuth, Al Khatib, and Stein* [Website][PDF]	ECPE-2D: Emotion-Cause Pair Extraction based on Joint Two-Dimensional Representation, Interaction and Prediction *Ding, Xia, and Yu* [Website][PDF]	Effective Inter-Clause Modeling for End-to-End Emotion-Cause Pair Extraction *Wei, Zhao, and Mao* [Website][PDF]	Embarrassingly Simple Unsupervised Aspect Extraction *Tulkens and Cranenburgh* [Website][PDF]	Enhancing Cross-target Stance Detection with Transferable Semantic-Emotion Knowledge *Zhang, Yang, Li, Ye, Xu, and Dai* [Website][PDF]
	KinGDOM: Knowledge-Guided DOMain Adaptation for Sentiment Analysis *Ghosal, Hazarika, Roy, Majumder, Mihalcea, and Poria* [Website][PDF]	Modelling Context and Syntactical Features for Aspect-based Sentiment Analysis *Phan and Ogunbona* [Website][PDF]	Parallel Data Augmentation for Formality Style Transfer *Zhang, Ge, and SUN* [Website][PDF]	Relational Graph Attention Network for Aspect-based Sentiment Analysis *Wang, Shen, Yang, Quan, and Wang* [Website][PDF]	SpanMlt: A Span-based Multi-Task Learning Framework for Pairwise Aspect and Opinion Terms Extraction *Zhao, Huang, Zhang, Lu, and* [Website][PDF]
	Syntax-Aware Opinion Role Labeling with Dependency Graph Convolutional Networks *Zhang, Zhang, Wang, Li, and Zhang* [Website][PDF]	[TACL] Target-Guided Structured Attention Network for Target-dependent Sentiment Analysis *Zhang, Chen, Liu, He, and Leung* [Website][PDF]	Towards Better Non-Tree Argument Mining: Proposition-Level Biaffine Parsing with Task-Specific Parameterization *Morio, Ozaki, Morishita, Koreeda, and Yanai* [Website][PDF]		
Track F *Student Research Workshop* Abstracts	uBLEU: Uncertainty-Aware Automatic Evaluation Method for Open-Domain Dialogue Systems *Yuma, Yoshinaga, and Toyoda* [Website][PDF]	To compress or not to compress? A Finite-State approach to Nen verbal morphology *Muradoglu, Evans, and Suominen* [Website][PDF]	AraDIC: Arabic Document Classification Using Image-Based Character Embeddings and Class-Balanced Loss *Daif, Kitada, and Iyatomi* [Website][PDF]	Self-Attention is Not Only a Weight: Analyzing BERT with Vector Norms *Kobayashi, Kuribayashi, Yokoi, and Inui* [Website]	
Track G *Tagging, Chunking and Parsing-1* Abstracts	[TACL] A Graph-based Model for Joint Chinese Word Segmentation and Dependency Parsing *Yan, Qiu, and Huang* [Website][PDF]	A Span-based Linearization for Constituent Trees *Wei, Wu, and Lan* [Website][PDF]	An Empirical Comparison of Unsupervised Constituency Parsing Methods *Li, Cao, Cai, Jiang, and Tu* [Website][PDF]	Efficient Constituency Parsing by Pointing *Nguyen, Nguyen, Joty, and Li* [Website][PDF]	Efficient Second-Order TreeCRF for Neural Dependency Parsing *Zhang, Li, and Zhang* [Website][PDF]

Representations of Syntax [MASK] Useful: Effects of Constituency and Dependency Structure in Recursive LSTMs *Lepori, Linzen, and McCoy* [Website] [PDF]	Structure-Level Knowledge Distillation For Multilingual Sequence Labeling *Wang, Jiang, Bach, Wang, Huang, and Tu* [Website] [PDF]

Session 6A Details

Session 6A: Ethics and NLP-1

Gender Bias in Multilingual Embeddings and Cross-Lingual Transfer [Website][PDF]
Jieyu Zhao, Subhabrata Mukherjee, saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah
5:00–6:00

Multilingual representations embed words from many languages into a single semantic space such that words with similar meanings are close to each other regardless of the language. These embeddings have been widely used in various settings, such as cross-lingual transfer, where a natural language processing (NLP) model trained on one language is deployed to another language. While the cross-lingual transfer techniques are powerful, they carry gender bias from the source to target languages. In this paper, we study gender bias in multilingual embeddings and how it affects transfer learning for NLP applications. We create a multilingual dataset for bias analysis and propose several ways for quantifying bias in multilingual representations from both the intrinsic and extrinsic perspectives. Experimental results show that the magnitude of bias in the multilingual representations changes differently when we align the embeddings to different target spaces and that the alignment direction can also have an influence on the bias in transfer learning. We further provide recommendations for using the multilingual word representations for downstream tasks.

Give Me Convenience and Give Her Death: Who Should Decide What Uses of NLP are Appropriate, and on What Basis? [Website][PDF]
kobi leins kobi, Jey Han Lau, and Timothy Baldwin 5:00–6:00

As part of growing NLP capabilities, coupled with an awareness of the ethical dimensions of research, questions have been raised about whether particular datasets and tasks should be deemed off-limits for NLP research. We examine this question with respect to a paper on automatic legal sentencing from EMNLP 2019 which was a source of some debate, in asking whether the paper should have been allowed to be published, who should have been charged with making such a decision, and on what basis. We focus in particular on the role of data statements in ethically assessing research, but also discuss the topic of dual use, and examine the outcomes of similar debates in other scientific disciplines.

Is Your Classifier Actually Biased? Measuring Fairness under Uncertainty with Bernstein Bounds [Website][PDF]
Kawin Ethayarajh 5:00–6:00

Most NLP datasets are not annotated with protected attributes such as gender, making it difficult to measure classification bias using standard measures of fairness (e.g., equal opportunity). However, manually annotating a large dataset with a protected attribute is slow and expensive. Instead of annotating all the examples, can we annotate a subset of them and use that sample to estimate the bias? While it is possible to do so, the smaller this annotated sample is, the less certain we are that the estimate is close to the true bias. In this work, we propose using Bernstein bounds to represent this uncertainty about the bias estimate as a confidence interval. We provide empirical evidence that a 95% confidence interval derived this way consistently bounds the true bias. In quantifying this uncertainty, our method, which we call Bernstein-bounded unfairness, helps prevent classifiers from being deemed biased or unbiased when there is insufficient evidence to make either claim. Our findings suggest that the datasets currently used to measure specific biases are too small to conclusively identify bias except in the most egregious cases. For example, consider a co-reference resolution system that is 5% more accurate on gender-stereotypical sentences – to claim it is biased with 95% confidence, we need a bias-specific dataset that is 3.8 times larger than WinoBias, the largest available.

It's Morphin' Time! Combating Linguistic Discrimination with Inflectional Perturbations [Website][PDF]
Samson Tan, Shafiq Joty, Min-Yen Kan, and Richard Socher 5:00–6:00

Training on only perfect Standard English corpora predisposes pre-trained neural networks to discriminate against minorities from non-standard linguistic backgrounds (e.g., African American Vernacular English, Colloquial Singapore English, etc.). We perturb the inflectional morphology of words to craft plausible and semantically similar adversarial examples that expose these biases in popular NLP models, e.g., BERT and Transformer, and show that adversarially fine-tuning them for a single epoch significantly improves robustness without sacrificing performance on clean data.

Mitigating Gender Bias Amplification in Distribution by Posterior Regularization [Website][PDF]
Shengyu Jia, Tao Meng, Jieyu Zhao, and Kai-Wei Chang 5:00–6:00

Advanced machine learning techniques have boosted the performance of natural language processing. Nevertheless, recent studies, e.g., (CITATION) show that these techniques inadvertently capture the societal bias hidden in the corpus and further amplify it. However, their analysis is conducted only on models' top predictions. In this paper, we investigate the gender bias amplification issue from the distribution perspective and demonstrate that the bias is amplified in the view of predicted probability distribution over labels. We further propose a bias mitigation approach based on posterior regularization. With little performance loss, our method can almost remove the bias amplification in the distribution. Our study sheds the light on understanding the bias amplification.

Towards Understanding Gender Bias in Relation Extraction [Website][PDF]
Andrew Gaut, Tony Sun, Shirlyn Tang, Yuxin Huang, Jing Qian, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang 5:00–6:00

Recent developments in Neural Relation Extraction (NRE) have made significant strides towards Automated Knowledge Base Construction. While much attention has been dedicated towards improvements in accuracy, there have been no attempts in the literature to evaluate social biases exhibited in NRE systems. In this paper, we create WikiGenderBias, a distantly supervised dataset composed of over 45,000 sentences including a 10% human annotated test set for the purpose of analyzing gender bias in relation extraction systems. We find that when extracting spouse-of and hypernym (i.e., occupation) relations, an NRE system performs differently when the gender of the target entity is different. However, such disparity does not appear when extracting relations such as birthDate or birthPlace. We also analyze how existing bias mitigation techniques, such as name anonymization, word embedding debiasing, and data augmentation affect the NRE system in terms of maintaining the test performance and reducing biases. Unfortunately, due to NRE models rely heavily on surface level cues, we find that existing bias mitigation approaches have a negative effect on NRE. Our analysis lays groundwork for future quantifying and mitigating bias in NRE.

Session 6A: Machine Learning for NLP-5

A Probabilistic Generative Model for Typographical Analysis of Early Modern Printing [Website][PDF]
Kartik Goyal, Chris Dyer, Christopher Warren, Maxwell G'Sell, and Taylor Berg-Kirkpatrick 5:00–6:00

We propose a deep and interpretable probabilistic generative model to analyze glyph shapes in printed Early Modern documents. We focus on clustering extracted glyph images into underlying templates in the presence of multiple confounding sources of variance. Our approach introduces a neural editor model that first generates well-understood printing phenomena like spatial perturbations from template parameters via interpertable latent variables, and then modifies the result by generating a non-interpretable latent vector responsible for inking variations, jitter, noise from the archiving process, and other unforeseen phenomena associated with Early Modern printing. Critically, by introducing an inference network whose input is restricted to the visual residual between the observation and the interpretably-modified template, we are able to control and isolate what the vector-valued latent variable captures. We show that our approach outperforms rigid interpretable clustering baselines (c.f. Ocular) and overly-flexible deep generative models (VAE) alike on the task of completely unsupervised discovery of typefaces in mixed-fonts documents.

Attentive Pooling with Learnable Norms for Text Representation [Website][PDF]
Chuhan Wu, Fangzhao Wu, Tao Qi, Xiaohui Cui, and Yongfeng Huang 5:00–6:00

Pooling is an important technique for learning text representations in many neural NLP models. In conventional pooling methods such as average, max and attentive pooling, text representations are weighted summations of the L1 or L∞ norm of input features. However, their pooling norms are always fixed and may not be optimal for learning accurate text representations in different tasks. In addition, in many popular pooling methods such as max and attentive pooling some features may be over-emphasized, while other useful ones are not fully exploited. In this paper, we propose an Attentive Pooling with Learnable Norms (APLN) approach for text representation. Different from existing pooling methods that use a fixed pooling norm, we propose to learn the norm in an end-to-end manner to automatically find the optimal ones for text representation in different tasks. In addition, we propose two methods to ensure the numerical stability of the model training. The first one is scale limiting, which re-scales the input to ensure non-negativity and alleviate the risk of exponential explosion. The second one is re-formulation, which decomposes the exponent operation to avoid computing the real-valued powers of the input and further accelerate the pooling operation. Experimental results on four benchmark datasets show that our approach can effectively improve the performance of attentive pooling.

Estimating the influence of auxiliary tasks for multi-task learning of sequence tagging tasks [Website][PDF]
Fynn Schröder and Chris Biemann 5:00–6:00

Multi-task learning (MTL) and transfer learning (TL) are techniques to overcome the issue of data scarcity when training state-of-the-art neural networks. However, finding beneficial auxiliary datasets for MTL or TL is a time- and resource-consuming trial-and-error approach. We propose new methods to automatically assess the similarity of sequence tagging datasets to identify beneficial auxiliary data for MTL or TL setups. Our methods can compute the similarity between any two sequence tagging datasets, i.e. they do not need to be annotated with the same tagset or multiple labels in parallel. Additionally, our methods take tokens and their labels into account, which is more robust than only using either of them as an information source, as conducted in prior work. We empirically show that our similarity measures correlate with the change in test score of neural networks that use the auxiliary dataset for MTL to increase the main task performance. We provide an efficient, open-source implementation.

How Does Selective Mechanism Improve Self-Attention Networks? [Website][PDF]
Xinwei Geng, Longyue Wang, Xing Wang, Bing Qin, Ting Liu, and Zhaopeng Tu 5:00–6:00

Self-attention networks (SANs) with selective mechanism has produced substantial improvements in various NLP tasks by concentrating on a subset of input words. However, the underlying reasons for their strong performance have not been well explained. In this paper, we bridge the gap by assessing the strengths of selective SANs (SSANs), which are implemented with a flexible and universal Gumbel-Softmax. Experimental results on several representative NLP tasks, including natural language inference, semantic role labelling, and machine translation, show that SSANs consistently outperform the standard SANs. Through well-designed probing experiments, we empirically validate that the improvement of SSANs can be attributed in part to mitigating two commonly-cited weaknesses of SANs: word order encoding and structure modeling. Specifically, the selective mechanism improves SANs by paying more attention to content words that contribute to the meaning of the sentence.

Improving Transformer Models by Reordering their Sublayers [Website][PDF]
Ofir Press, Noah A. Smith, and Omer Levy 5:00–6:00

Multilayer transformer networks consist of interleaved self-attention and feedforward sublayers. Could ordering the sublayers in a different pattern lead to better performance? We generate randomly ordered transformers and train them with the language modeling objective. We observe that some of these models are able to achieve better performance than the interleaved baseline, and that those successful variants tend to have more self-attention at the bottom and more feedforward sublayers at the top. We propose a new transformer pattern that adheres to this property, the sandwich transformer, and show that it improves perplexity on multiple word-level and character-level language modeling benchmarks, at no cost in parameters, memory, or training time. However, the sandwich reordering pattern does not guarantee performance gains across every task, as we demonstrate on machine translation models. Instead, we suggest that further exploration of task-specific sublayer reorderings is needed in order to unlock additional gains.

Single Model Ensemble using Pseudo-Tags and Distinct Vectors [Website][PDF]
Ryosuke Kuwabara, Jun Suzuki, and Hideki Nakayama 5:00–6:00

Model ensemble techniques often increase task performance in neural networks; however, they require increased time, memory, and management effort. In this study, we propose a novel method that replicates the effects of a model ensemble with a single model. Our approach creates K-virtual models within a single parameter space using K-distinct pseudo-tags and K-distinct vectors. Experiments on text classification and sequence labeling tasks on several datasets demonstrate that our method emulates or outperforms a traditional model ensemble with 1/K-times fewer parameters.

Zero-shot Text Classification via Reinforced Self-training [Website][PDF]
Zhiquan Ye, Yuxia Geng, Jiaoyan Chen, Jingmin Chen, Xiaoxiao Xu, SuHang Zheng, Feng Wang, Jun Zhang, and Huajun Chen 5:00–6:00

Zero-shot learning has been a tough problem since no labeled data is available for unseen classes during training, especially for classes with low similarity. In this situation, transferring from seen classes to unseen classes is extremely hard. To tackle this problem, in this paper we propose a self-training based method to efficiently leverage unlabeled data. Traditional self-training methods use fixed heuristics to select instances from unlabeled data, whose performance varies among different datasets. We propose a reinforcement learning framework to learn data selection strategy automatically and provide more reliable selection. Experimental results on both benchmarks and a real-world e-commerce dataset show that our approach significantly outperforms previous methods in zero-shot text classification

Session 6A: Machine Translation-7

A Novel Graph-based Multi-modal Fusion Encoder for Neural Machine Translation [Website][PDF]
Yongjing Yin, Fandong Meng, Jinsong Su, Chulun Zhou, Zhengyuan Yang, Jie Zhou, and Jiebo Luo 5:00–6:00

Multi-modal neural machine translation (NMT) aims to translate source sentences into a target language paired with images. However, dominant multi-modal NMT models do not fully exploit fine-grained semantic correspondences between semantic units of different modalities, which have potential to refine multi-modal representation learning. To deal with this issue, in this paper, we propose a novel graph-based multi-modal fusion encoder for NMT. Specifically, we first represent the input sentence and image using a unified multi-modal graph, which captures various semantic relationships between multi-modal semantic units (words and visual objects). We then stack multiple graph-based multi-modal fusion layers that iteratively perform semantic interactions to learn node representations. Finally, these representations provide an attention-based context vector for the decoder. We evaluate our proposed encoder on the Multi30K datasets. Experimental results and in-depth analysis show the superiority of our multi-modal NMT model.

A Relaxed Matching Procedure for Unsupervised BLI [Website][PDF]
Xu Zhao, Zihao Wang, Yong Zhang, and Hao Wu 5:00–6:00

Recently unsupervised Bilingual Lexicon Induction(BLI) without any parallel corpus has attracted much research interest. One of the crucial parts in methods for the BLI task is the matching procedure. Previous works impose a too strong constraint on the matching and lead to many counterintuitive translation pairings. Thus We propose a relaxed matching procedure to find a more precise matching between two languages. We also find that aligning source and target language embedding space bidirectionally will bring significant improvement. We follow the previous iterative framework to conduct experiments. Results on standard benchmark demonstrate the effectiveness of our proposed method, which substantially outperforms previous unsupervised methods.

Dynamic Programming Encoding for Subword Segmentation in Neural Machine Translation [Website][PDF]
Xuanli He, Gholamreza Haffari, and Mohammad Norouzi 5:00–6:00

This paper introduces Dynamic Programming Encoding (DPE), a new segmentation algorithm for tokenizing sentences into subword units. We view the subword segmentation of output sentences as a latent variable that should be marginalized out for learning and inference. A mixed character-subword transformer is proposed, which enables exact log marginal likelihood estimation and exact MAP inference to find target segmentations with maximum posterior probability. DPE uses a lightweight mixed character-subword transformer as a means of pre-processing parallel data to segment output sentences using dynamic programming. Empirical results on machine translation suggest that DPE is effective for segmenting output sentences and can be combined with BPE dropout for stochastic segmentation of source sentences. DPE achieves an average improvement of 0.9 BLEU over BPE (Sennrich et al., 2016) and an average improvement of 0.55 BLEU over BPE dropout (Provilkov et al., 2019) on several WMT datasets including English <=> (German, Romanian, Estonian, Finnish, Hungarian).

Geometry-aware domain adaptation for unsupervised alignment of word embeddings [Website][PDF]
Pratik Jawanpuria, Mayank Meghwanshi, and Bamdev Mishra 5:00–6:00

We propose a novel manifold based geometric approach for learning unsupervised alignment of word embeddings between the source and the target languages. Our approach formulates the alignment learning problem as a domain adaptation problem over the manifold of doubly stochastic matrices. This viewpoint arises from the aim to align the second order information of the two language spaces. The rich geometry of the doubly stochastic manifold allows to employ efficient Riemannian conjugate gradient algorithm for the proposed formulation. Empirically, the proposed approach outperforms state-of-the-art optimal transport based approach on the bilingual lexicon induction task across several language pairs. The performance improvement is more significant for distant language pairs.

Learning to Recover from Multi-Modality Errors for Non-Autoregressive Neural Machine Translation [Website][PDF]
Qiu Ran, Yankai Lin, Peng Li, and Jie Zhou 5:00–6:00

Non-autoregressive neural machine translation (NAT) predicts the entire target sequence simultaneously and significantly accelerates inference process. However, NAT discards the dependency information in a sentence, and thus inevitably suffers from the multi-modality problem: the target tokens may be provided by different possible translations, often causing token repetitions or missing. To alleviate this problem, we propose a novel semi-autoregressive model RecoverSAT in this work, which generates a translation as a sequence of segments. The segments are generated simultaneously while each segment is predicted token-by-token. By dynamically determining segment length and deleting repetitive segments, RecoverSAT is capable of recovering from repetitive and missing token errors. Experimental results on three widely-used benchmark datasets show that our proposed model achieves more than 4 times speedup while maintaining comparable performance compared with the corresponding autoregressive model.

On the Inference Calibration of Neural Machine Translation [Website][PDF]
Shuo Wang, Zhaopeng Tu, Shuming Shi, and Yang Liu 5:00–6:00

Confidence calibration, which aims to make model predictions equal to the true correctness measures, is important for neural machine translation (NMT) because it is able to offer useful indicators of translation errors in the generated output. While prior studies have shown that NMT models trained with label smoothing are well-calibrated on the

ground-truth training data, we find that miscalibration still remains a severe challenge for NMT during inference due to the discrepancy between training and inference. By carefully designing experiments on three language pairs, our work provides in-depth analyses of the correlation between calibration and translation performance as well as linguistic properties of miscalibration and reports a number of interesting findings that might help humans better analyze, understand and improve NMT models. Based on these observations, we further propose a new graduated label smoothing method that can improve both inference calibration and translation performance.

[CL] Unsupervised Word Translation with Adversarial Autoencoder [Website] [PDF]
Tasnim Mohiuddin and Shafiq Joty 5:00–6:00

Crosslingual word embeddings learned from monolingual embeddings have a crucial role in many downstream tasks, ranging from machine translation to transfer learning. Adversarial training has shown impressive success in learning crosslingual embeddings and the associated word translation task without any parallel data by mapping monolingual embeddings to a shared space. However, recent work has shown superior performance for non-adversarial methods in more challenging language pairs. In this article, we investigate adversarial autoencoder for unsupervised word translation and propose two novel extensions to it that yield more stable training and improved results. Our method includes regularization terms to enforce cycle consistency and input reconstruction, and puts the target encoders as an adversary against the corresponding discriminator. We use two types of refinement procedures sequentially after obtaining the trained encoders and mappings from the adversarial training, namely, refinement with Procrustes solution and refinement with symmetric re-weighting. Extensive experimentations with high- and low-resource languages from two different data sets show that our method achieves better performance than existing adversarial and non-adversarial approaches and is also competitive with the supervised system. Along with performing comprehensive ablation studies to understand the contribution of different components of our adversarial model, we also conduct a thorough analysis of the refinement procedures to understand their effects.

Session 6A: NLP Applications-5

Camouflaged Chinese Spam Content Detection with Semi-supervised Generative Active Learning [Website][PDF]
Zhuoren Jiang, Zhe Gao, Yu Duan, Yangyang Kang, Changlong Sun, Qiong Zhang, and Xiaozhong Liu
5:00–6:00

We propose a Semi-supervIsed GeNerative Active Learning (SIGNAL) model to address the imbalance, efficiency, and text camouflage problems of Chinese text spam detection task. A "self-diversity" criterion is proposed for measuring the "worthiness" of a candidate for annotation. A semi-supervised variational autoencoder with masked attention learning approach and a character variation graph-enhanced augmentation procedure are proposed for data augmentation. The preliminary experiment demonstrates the proposed SIGNAL model is not only sensitive to spam sample selection, but also can improve the performance of a series of conventional active learning models for Chinese spam detection task. To the best of our knowledge, this is the first work to integrate active learning and semi-supervised generative learning for text spam detection.

Distinguish Confusing Law Articles for Legal Judgment Prediction [Website][PDF]
Nuo Xu, Pinghui Wang, Long Chen, Li Pan, Xiaoyan Wang, and Junzhou Zhao 5:00–6:00

Legal Judgement Prediction (LJP) is the task of automatically predicting a law case's judgment results given a text describing the case's facts, which has great prospects in judicial assistance systems and handy services for the public. In practice, confusing charges are often presented, because law cases applicable to similar law articles are easily misjudged. To address this issue, existing work relies heavily on domain experts, which hinders its application in different law systems. In this paper, we present an end-to-end model, LADAN, to solve the task of LJP. To distinguish confusing charges, we propose a novel graph neural network, GDL, to automatically learn subtle differences between confusing law articles, and also design a novel attention mechanism that fully exploits the learned differences to attentively extract effective discriminative features from fact descriptions. Experiments conducted on real-world datasets demonstrate the superiority of our LADAN.

Hiring Now: A Skill-Aware Multi-Attention Model for Job Posting Generation [Website][PDF]
Liting Liu, Jie Liu, Wenzheng Zhang, Ziming Chi, Wenxuan Shi, and Yalou Huang 5:00–6:00

Writing a good job posting is a critical step in the recruiting process, but the task is often more difficult than many people think. It is challenging to specify the level of education, experience, relevant skills per the company information and job description. To this end, we propose a novel task of Job Posting Generation (JPG) which is cast as a conditional text generation problem to generate job requirements according to the job descriptions. To deal with this task, we devise a data-driven global Skill-Aware Multi-Attention generation model, named SAMA. Specifically, to model the complex mapping relationships between input and output, we design a hierarchical decoder that we first label the job description with multiple skills, then we generate a complete text guided by the skill labels. At the same time, to exploit the prior knowledge about the skills, we further construct a skill knowledge graph to capture the global prior knowledge of skills and refine the generated results. The proposed approach is evaluated on real-world job posting data. Experimental results clearly demonstrate the effectiveness of the proposed method.

HyperCore: Hyperbolic and Co-graph Representation for Automatic ICD Coding [Website][PDF]
Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Shengping Liu, and Weifeng Chong 5:00–6:00

The International Classification of Diseases (ICD) provides a standardized way for classifying diseases, which endows each disease with a unique code. ICD coding aims to assign proper ICD codes to a medical record. Since manual coding is very laborious and prone to errors, many methods have been proposed for the automatic ICD coding task. However, most of existing methods independently predict each code, ignoring two important characteristics: Code Hierarchy and Code Co-occurrence. In this paper, we propose a Hyperbolic and Co-graph Representation method (HyperCore) to address the above problem. Specifically, we propose a hyperbolic representation method to leverage the code hierarchy. Moreover, we propose a graph convolutional network to utilize the code co-occurrence. Experimental results on two widely used datasets demonstrate that our proposed model outperforms previous state-of-the-art methods.

Hyperbolic Capsule Networks for Multi-Label Classification [Website][PDF]
Boli Chen, Xin Huang, Lin Xiao, and Liping Jing 5:00–6:00

Although deep neural networks are effective at extracting high-level features, classification methods usually encode an input into a vector representation via simple feature aggregation operations (e.g. pooling). Such operations limit the performance. For instance, a multi-label document may contain several concepts. In this case, one vector can not sufficiently capture its salient and discriminative content. Thus, we propose Hyperbolic Capsule Networks (HyperCaps) for Multi-Label Classification (MLC), which have two merits. First, hyperbolic capsules are designed to capture fine-grained document information for each label, which has the ability to characterize complicated structures among labels and documents. Second, Hyperbolic Dynamic Routing (HDR) is introduced to aggregate hyperbolic capsules in a label-aware manner, so that the label-level discriminative information can be preserved along the depth of neural networks. To efficiently handle large-scale MLC datasets, we additionally present a new routing method to adaptively adjust the capsule number during routing. Extensive experiments are conducted on four benchmark datasets. Compared with the state-of-the-art methods, HyperCaps significantly improves the performance of MLC especially on tail labels.

Improving Segmentation for Technical Support Problems [Website][PDF]
Kushal Chauhan and Abhirut Gupta 5:00–6:00

Technical support problems are often long and complex. They typically contain user descriptions of the problem, the setup, and steps for attempted resolution. Often they also contain various non-natural language text elements like outputs of commands, snippets of code, error messages or stack traces. These elements contain potentially crucial information for problem resolution. However, they cannot be correctly parsed by tools designed for natural language. In this paper, we address the problem of segmentation for technical support questions. We formulate the problem as a sequence labelling task, and study the performance of state of the art approaches. We compare this against an intuitive contextual sentence-level classification baseline, and a state of the art supervised text-segmentation approach. We also introduce a novel component of combining contextual embeddings from multiple language models pre-trained on different data sources, which achieves a marked improvement over using embeddings from a single pre-trained language model. Finally, we also demonstrate the usefulness of such segmentation with improvements on the downstream task of answer retrieval.

MOOCCube: A Large-scale Data Repository for NLP Applications in MOOCs [Website] [PDF]
Jifan Yu, Gan Luo, Tong Xiao, Qingyang Zhong, Yuquan Wang, wenzheng feng wenzheng, Junyi Luo, Chenyu Wang, Lei Hou, Juanzi Li, Zhiyuan Liu, and Jie Tang 5:00–6:00

The prosperity of Massive Open Online Courses (MOOCs) provides fodder for many NLP and AI research for education applications, e.g., course concept extraction, prerequisite relation discovery, etc. However, the publicly available datasets of MOOC are limited in size with few types of data, which hinders advanced models and novel attempts in related topics. Therefore, we present MOOCCube, a large-scale data repository of over 700 MOOC courses, 100k concepts, 8 million student behaviors with an external resource. Moreover, we conduct a prerequisite discovery task as an example application to show the potential of MOOCCube in facilitating relevant research. The data repository is now available at http://moocdata.cn/data/MOOCCube.

Towards Interpretable Clinical Diagnosis with Bayesian Network Ensembles Stacked on Entity-Aware CNNs [Website] [PDF]
Jun Chen, Xiaoya Dai, Quan Yuan, Chao Lu, and Haifeng Huang 5:00–6:00

The automatic text-based diagnosis remains a challenging task for clinical use because it requires appropriate balance between accuracy and interpretability. In this paper, we attempt to propose a solution by introducing a novel framework that stacks Bayesian Network Ensembles on top of Entity-Aware Convolutional Neural Networks (CNN) towards building an accurate yet interpretable diagnosis system. The proposed framework takes advantage of the high accuracy and generality of deep neural networks as well as the interpretability of Bayesian Networks, which is critical for AI-empowered healthcare. The evaluation conducted on the real Electronic Medical Record (EMR) documents from hospitals and annotated by professional doctors proves that, the proposed framework outperforms the previous automatic diagnosis methods in accuracy performance and the diagnosis explanation of the framework is reasonable.

Session 6A: Sentiment Analysis, Stylistic Analysis, and Argument Mining-1

Analyzing the Persuasive Effect of Style in News Editorial Argumentation [Website][PDF]
Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein 5:00–6:00

News editorials argue about political issues in order to challenge or reinforce the stance of readers with different ideologies. Previous research has investigated such persuasive effects for argumentative content. In contrast, this paper studies how important the style of news editorials is to achieve persuasion. To this end, we first compare content- and style-oriented classifiers on editorials from the liberal NYTimes with ideology-specific effect annotations. We find that conservative readers are resistant to NYTimes style, but on liberals, style even has more impact than content. Focusing on liberals, we then cluster the leads, bodies, and endings of editorials, in order to learn about writing style patterns of effective argumentation.

ECPE-2D: Emotion-Cause Pair Extraction based on Joint Two-Dimensional Representation, Interaction and Prediction [Website][PDF]
Zixiang Ding, Rui Xia, and Jianfei Yu 5:00–6:00

In recent years, a new interesting task, called emotion-cause pair extraction (ECPE), has emerged in the area of text emotion analysis. It aims at extracting the potential pairs of emotions and their corresponding causes in a document. To solve this task, the existing research employed a two-step framework, which first extracts individual emotion set and cause set, and then pair the corresponding emotions and causes. However, such a pipeline of two steps contains some inherent flaws: 1) the modeling does not aim at extracting the final emotion-cause pair directly; 2) the errors from the first step will affect the performance of the second step. To address these shortcomings, in this paper we propose a new end-to-end approach, called ECPE-Two-Dimensional (ECPE-2D), to represent the emotion-cause pairs by a 2D representation scheme. A 2D transformer module and two variants, window-constrained and cross-road 2D transformers, are further proposed to model the interactions of different emotion-cause pairs. The 2D representation, interaction, and prediction are integrated into a joint framework. In addition to the advantages of joint modeling, the experimental results on the benchmark emotion cause corpus show that our approach improves the F1 score of the state-of-the-art from 61.28% to 68.89%.

Effective Inter-Clause Modeling for End-to-End Emotion-Cause Pair Extraction [Website][PDF]
Penghui Wei, Jiahao Zhao, and Wenji Mao 5:00–6:00

Emotion-cause pair extraction aims to extract all emotion clauses coupled with their cause clauses from a given document. Previous work employs two-step approaches, in which the first step extracts emotion clauses and cause clauses separately, and the second step trains a classifier to filter out negative pairs. However, such pipeline-style system for emotion-cause pair extraction is suboptimal because it suffers from error propagation and the two steps may not adapt to each other well. In this paper, we tackle emotion-cause pair extraction from a ranking perspective, i.e., ranking clause pair candidates in a document, and propose a one-step neural approach which emphasizes inter-clause modeling to perform end-to-end extraction. It models the interrelations between the clauses in a document to learn clause representations with graph attention, and enhances clause pair representations with kernel-based relative position embedding for effective ranking. Experimental results show that our approach significantly outperforms the current two-step systems, especially in the condition of extracting multiple pairs in one document.

Embarrassingly Simple Unsupervised Aspect Extraction [Website][PDF]
Stéphan Tulkens and Andreas van Cranenburgh 5:00–6:00

We present a simple but effective method for aspect identification in sentiment analysis. Our unsupervised method only requires word embeddings and a POS tagger, and is therefore straightforward to apply to new domains and languages. We introduce Contrastive Attention (CAt), a novel single-head attention mechanism based on an RBF kernel, which gives a considerable boost in performance and makes the model interpretable. Previous work relied on syntactic features and complex neural models. We show that given the simplicity of current benchmark datasets for aspect extraction, such complex models are not needed. The code to reproduce the experiments reported in this paper is available at https://github.com/clips/cat.

Enhancing Cross-target Stance Detection with Transferable Semantic-Emotion Knowledge [Website][PDF]
Bowen Zhang, Min Yang, Xutao Li, Yunming Ye, Xiaofei Xu, and Kuai Dai 5:00–6:00

Stance detection is an important task, which aims to classify the attitude of an opinionated text towards a given target. Remarkable success has been achieved when sufficient labeled training data is available. However, annotating sufficient data is labor-intensive, which establishes significant barriers for generalizing the stance classifier to the data with new targets. In this paper, we proposed a Semantic-Emotion Knowledge Transferring (SEKT) model for cross-target stance detection, which uses the external knowledge (semantic and emotion lexicons) as a bridge to enable knowledge transfer across different targets. Specifically, a semantic-emotion heterogeneous graph is constructed from external semantic and emotion lexicons, which is then fed into a graph convolutional network to learn multi-hop semantic connections between words and emotion tags. Then, the learned semantic-emotion graph representation, which serves as prior knowledge bridging the gap between the source and target domains, is fully integrated into the bidirectional long short-term memory (BiLSTM) stance classifier by adding a novel knowledge-aware memory unit to the BiLSTM cell. Extensive experiments on a large real-world dataset demonstrate the superiority of SEKT against the state-of-the-art baseline methods.

KinGDOM: Knowledge-Guided DOMain Adaptation for Sentiment Analysis [Website][PDF]
Deepanway Ghosal, Devamanyu Hazarika, Abhinaba Roy, Navonil Majumder, Rada Mihalcea, and Soujanya Poria 5:00–6:00

Cross-domain sentiment analysis has received significant attention in recent years, prompted by the need to combat the domain gap between different applications that make use of sentiment analysis. In this paper, we take a novel perspective on this task by exploring the role of external commonsense knowledge. We introduce a new framework, KinGDOM, which utilizes the ConceptNet knowledge graph to enrich the semantics of a document by providing both domain-specific and domain-general background concepts. These concepts are learned by training a graph convolutional autoencoder that leverages inter-domain concepts in a domain-invariant manner. Conditioning a popular domain-adversarial baseline method with these learned concepts helps improve its performance over state-of-the-art approaches, demonstrating the efficacy of our proposed framework.

Modelling Context and Syntactical Features for Aspect-based Sentiment Analysis [Website][PDF]
Minh Hieu Phan and Philip O. Ogunbona 5:00–6:00

The aspect-based sentiment analysis (ABSA) consists of two conceptual tasks, namely an aspect extraction and an aspect sentiment classification. Rather than considering the tasks separately, we build an end-to-end ABSA solution. Previous works in ABSA tasks did not fully leverage the importance of syntactical information. Hence, the aspect extraction model often failed to detect the boundaries of multi-word aspect terms. On the other hand, the aspect sentiment classifier was unable to account for the syntactical correlation between aspect terms and the context words. This paper explores the grammatical aspect of the sentence and employs the self-attention mechanism for syntactical learning. We combine part-of-speech embeddings, dependency-based embeddings and contextualized embeddings (e.g. BERT, RoBERTa) to enhance the performance of the aspect extractor. We also propose the syntactic relative distance to de-emphasize the adverse effects of unrelated words, having weak syntactic connection with the aspect terms. This increases the accuracy of the aspect sentiment classifier. Our solutions outperform the state-of-the-art models on SemEval-2014 dataset in both two subtasks.

Parallel Data Augmentation for Formality Style Transfer [Website][PDF]
Yi Zhang, Tao Ge, and Xu SUN 5:00–6:00

The main barrier to progress in the task of Formality Style Transfer is the inadequacy of training data. In this paper, we study how to augment parallel data and propose novel and simple data augmentation methods for this task to obtain useful sentence pairs with easily accessible models and systems. Experiments demonstrate that our augmented parallel data largely helps improve formality style transfer when it is used to pre-train the model, leading to the state-of-the-art results in the GYAFC benchmark dataset.

Relational Graph Attention Network for Aspect-based Sentiment Analysis [Website][PDF]
Kai Wang, Weizhou Shen, Yunyi Yang, Xiaojun Quan, and Rui Wang 5:00–6:00

Aspect-based sentiment analysis aims to determine the sentiment polarity towards a specific aspect in online reviews. Most recent efforts adopt attention-based neural network models to implicitly connect aspects with opinion words. However, due to the complexity of language and the existence of multiple aspects in a single sentence, these models often confuse the connections. In this paper, we address this problem by means of effective encoding of syntax information. Firstly, we define a unified aspect-oriented dependency tree structure rooted at a target aspect by reshaping and pruning an ordinary dependency parse tree. Then, we propose a relational graph attention network (R-GAT) to encode the new tree structure for sentiment prediction. Extensive experiments are conducted on the SemEval 2014 and Twitter datasets, and the experimental results confirm that the connections between aspects and opinion words can be better established with our approach, and the performance of the graph attention network (GAT) is significantly improved as a consequence.

SpanMlt: A Span-based Multi-Task Learning Framework for Pair-wise Aspect and Opinion Terms Extraction [Website][PDF]
He Zhao, Longtao Huang, Rong Zhang, Quan Lu, and hui xue hui 5:00–6:00

Aspect terms extraction and opinion terms extraction are two key problems of fine-grained Aspect Based Sentiment Analysis (ABSA). The aspect-opinion pairs can provide a global profile about a product or service for consumers and opinion mining systems. However, traditional methods can not directly output aspect-opinion pairs without given aspect terms or opinion terms. Although some recent co-extraction methods have been proposed to extract both terms jointly, they fail to extract them as pairs. To this end, this paper proposes an end-to-end method to solve the task of Pair-wise Aspect and Opinion Terms Extraction (PAOTE). Furthermore, this paper treats the problem from a perspective of joint term and relation extraction rather than under the sequence tagging formulation performed in most prior works. We propose a multi-task learning framework based on shared spans, where the terms are extracted under the supervision of span boundaries. Meanwhile, the pair-wise relations are jointly identified using the span representations. Extensive experiments show that our model consistently outperforms state-of-the-art methods.

Syntax-Aware Opinion Role Labeling with Dependency Graph Convolutional Networks [Website][PDF]
Bo Zhang, Yue Zhang, Rui Wang, Zhenghua Li, and Min Zhang 5:00–6:00

Opinion role labeling (ORL) is a fine-grained opinion analysis task and aims to answer "who expressed what kind of sentiment towards what?". Due to the scarcity of labeled data, ORL remains challenging for data-driven methods. In this work, we try to enhance neural ORL models with syntactic knowledge by comparing and integrating different representations. We also propose dependency graph convolutional networks (DEPGCN) to encode parser information at different processing levels. In order to compensate for parser inaccuracy and reduce error propagation, we introduce multi-task learning (MTL) to train the parser and the ORL model simultaneously. We verify our methods on the benchmark MPQA corpus. The experimental results show that syntactic information is highly valuable for ORL,

and our final MTL model effectively boosts the F1 score by 9.29 over the syntax-agnostic baseline. In addition, we find that the contributions from syntactic knowledge do not fully overlap with contextualized word representations (BERT). Our best model achieves 4.34 higher F1 score than the current state-ofthe-art.

[TACL] Target-Guided Structured Attention Network for Target-dependent Sentiment Analysis [Website][PDF]
Ji Zhang, Chengyao Chen, Pengfei Liu, Chao He, and Cane Wing-Ki Leung 5:00–6:00

Target-dependent sentiment analysis (TDSA) aims to classify the sentiment of a text towards a given target. The major challenge of this task lies in modeling the semantic relatedness between a target and its context sentence. This paper proposes a novel Target-Guided Structured Attention Network (TG-SAN), which captures target-related contexts for TDSA in a fine-to-coarse manner. Given a target and its context sentence, the proposed TG-SAN first identifies multiple semantic segments from the sentence using a target-guided structured attention mechanism. It then fuses the extracted segments based on their relatedness with the target for sentiment classification. We present comprehensive comparative experiments on three benchmarks with three major findings. Firstly, TG-SAN outperforms the state-of-the-art by up to 1.61% and 3.58% in terms of accuracy and Marco-F1 respectively. Secondly, it shows a strong advantage in determining the sentiment of a target when the context sentence contains multiple semantic segments. Lastly, the attention results produced by TG-SAN are highly interpretable as visualization results shown.

Towards Better Non-Tree Argument Mining: Proposition-Level Biaffine Parsing with Task-Specific Parameterization [Website][PDF]
Gaku Morio, Hiroaki Ozaki, Terufumi Morishita, Yuta Koreeda, and Kohsuke Yanai 5:00–6:00

State-of-the-art argument mining studies have advanced the techniques for predicting argument structures. However, the technology for capturing non-tree-structured arguments is still in its infancy. In this paper, we focus on non-tree argument mining with a neural model. We jointly predict proposition types and edges between propositions. Our proposed model incorporates (i) task-specific parameterization (TSP) that effectively encodes a sequence of propositions and (ii) a proposition-level biaffine attention (PLBA) that can predict a non-tree argument consisting of edges. Experimental results show that both TSP and PLBA boost edge prediction performance compared to baselines.

Session 6A: Student Research Workshop

uBLEU: Uncertainty-Aware Automatic Evaluation Method for Open-Domain Dialogue Systems [Website][PDF]
Tsuta Yuma, Naoki Yoshinaga, and Masashi Toyoda 5:00–6:00

Because open-domain dialogues allow diverse responses, basic reference-based metrics such as BLEU do not work well unless we prepare a massive reference set of high-quality responses for input utterances. To reduce this burden, a human-aided, uncertainty-aware metric, ΔBLEU, has been proposed; it embeds human judgment on the quality of reference outputs into the computation of multiple-reference BLEU. In this study, we instead propose a fully automatic, uncertainty-aware evaluation method for open-domain dialogue systems, *v*BLEU. This method first collects diverse reference responses from massive dialogue data and then annotates their quality judgments by using a neural network trained on automatically collected training data. Experimental results on massive Twitter data confirmed that *v*BLEU is comparable to ΔBLEU in terms of its correlation with human judgment and that the state of the art automatic evaluation method, RUBER, is improved by integrating *v*BLEU.

To compress or not to compress? A Finite-State approach to Nen verbal morphology [Website][PDF]
Saliha Muradoglu, Nicholas Evans, and Hanna Suominen 5:00–6:00

This paper describes the development of a verbal morphological parser for an under-resourced Papuan language, Nen. Nen verbal morphology is particularly complex, with a transitive verb taking up to 1,740 unique features. The structural properties exhibited by Nen verbs raises interesting choices for analysis. Here we compare two possible methods of analysis: 'Chunking' and decomposition. 'Chunking' refers to the concept of collating morphological segments into one, whereas the decomposition model follows a more classical linguistic approach. Both models are built using the Finite-State Transducer toolkit foma. The resultant architecture shows differences in size and structural clarity. While the 'Chunking' model is under half the size of the full de-composed counterpart, the decomposition displays higher structural order. In this paper, we describe the challenges encountered when modelling a language exhibiting distributed exponence and present the first morphological analyser for Nen, with an overall accuracy of 80.3%.

AraDIC: Arabic Document Classification Using Image-Based Character Embeddings and Class-Balanced Loss [Website][PDF]
Mahmoud Daif, Shunsuke Kitada, and Hitoshi Iyatomi 5:00–6:00

Classical and some deep learning techniques for Arabic text classification often depend on complex morphological analysis, word segmentation, and hand-crafted feature engineering. These could be eliminated by using character-level features. We propose a novel end-to-end Arabic document classification framework, Arabic document image-based classifier (AraDIC), inspired by the work on image-based character embeddings. AraDIC consists of an image-based character encoder and a classifier. They are trained in an end-to-end fashion using the class balanced loss to deal with the long-tailed data distribution problem. To evaluate the effectiveness of AraDIC, we created and published two datasets, the Arabic Wikipedia title (AWT) dataset and the Arabic poetry (AraP) dataset. To the best of our knowledge, this is the first image-based character embedding framework addressing the problem of Arabic text classification. We also present the first deep learning-based text classifier widely evaluated on modern standard Arabic, colloquial Arabic, and Classical Arabic. AraDIC shows performance improvement over classical and deep learning baselines by 12.29% and 23.05% for the micro and macro F-score, respectively.

Self-Attention is Not Only a Weight: Analyzing BERT with Vector Norms [Website]
Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui 5:00–6:00

Self-attention modules are essential building blocks of Transformer-based language models and hence are the subject of a large number of studies aiming to discover which linguistic capabilities these models possess (Rogers et al., 2020). Such studies are commonly conducted by analyzing correlations of attention weights with specific linguistic phenomena. In this paper, we show that attention weights alone are only one of two factors determining the output of self-attention modules and propose to incorporate the other factor, namely the norm of the transformed input vectors, into the analysis, as well. Our analysis of self-attention modules in BERT (Devlin et al., 2019) shows that the proposed method produces insights that better agree with linguistic intuitions than an analysis based on attention-weights alone. Our analysis further reveals that BERT controls the amount of the contribution from frequent informative and less informative tokens not by attention weights but via vector norms.

Session 6A Syntax: Tagging, Chunking and Parsing-1

[TACL] A Graph-based Model for Joint Chinese Word Segmentation and Dependency Parsing [Website][PDF]
Hang Yan, Xipeng Qiu, and Xuanjing Huang 5:00–6:00

Chinese word segmentation and dependency parsing are two fundamental tasks for Chinese natural language processing. The dependency parsing is defined on word-level. Therefore, word segmentation is the precondition of dependency parsing, which makes dependency parsing suffer from error propagation and unable to directly make use of the character-level pre-trained language model (such as BERT). In this paper, we propose a graph-based model to integrate Chinese word segmentation and dependency parsing. Different from previous transition-based joint models, our proposed model is more concise, which results in fewer efforts of feature engineering. Our graph-based joint model achieves better performance than previous joint models and state-of-the-art results in both Chinese word segmentation and dependency parsing. Besides, when BERT is combined, our model can substantially reduce the performance gap of dependency parsing between joint models and gold-segmented word-based models. Our code is publicly available at https://github.com/fastnlp/JointCwsParser.

A Span-based Linearization for Constituent Trees [Website][PDF]
Yang Wei, Yuanbin Wu, and Man Lan 5:00–6:00

We propose a novel linearization of a constituent tree, together with a new locally normalized model. For each split point in a sentence, our model computes the normalizer on all spans ending with that split point, and then predicts a tree span from them. Compared with global models, our model is fast and parallelizable. Different from previous local models, our linearization method is tied on the spans directly and considers more local features when performing span prediction, which is more interpretable and effective. Experiments on PTB (95.8 F1) and CTB (92.4 F1) show that our model significantly outperforms existing local models and efficiently achieves competitive results with global models.

An Empirical Comparison of Unsupervised Constituency Parsing Methods [Website][PDF]
Jun Li, Yifan Cao, Jiong Cai, Yong Jiang, and Kewei Tu 5:00–6:00

Unsupervised constituency parsing aims to learn a constituency parser from a training corpus without parse tree annotations. While many methods have been proposed to tackle the problem, including statistical and neural methods, their experimental results are often not directly comparable due to discrepancies in datasets, data preprocessing, lexicalization, and evaluation metrics. In this paper, we first examine experimental settings used in previous work and propose to standardize the settings for better comparability between methods. We then empirically compare several existing methods, including decade-old and newly proposed ones, under the standardized settings on English and Japanese, two languages with different branching tendencies. We find that recent models do not show a clear advantage over decade-old models in our experiments. We hope our work can provide new insights into existing methods and facilitate future empirical evaluation of unsupervised constituency parsing.

Efficient Constituency Parsing by Pointing [Website][PDF]
Thanh-Tung Nguyen, Xuan-Phi Nguyen, Shafiq Joty, and Xiaoli Li 5:00–6:00

We propose a novel constituency parsing model that casts the parsing problem into a series of pointing tasks. Specifically, our model estimates the likelihood of a span being a legitimate tree constituent via the pointing score corresponding to the boundary words of the span. Our parsing model supports efficient top-down decoding and our learning objective is able to enforce structural consistency without resorting to the expensive CKY inference. The experiments on the standard English Penn Treebank parsing task show that our method achieves 92.78 F1 without using pre-trained models, which is higher than all the existing methods with similar time complexity. Using pre-trained BERT, our model achieves 95.48 F1, which is competitive with the state-of-the-art while being faster. Our approach also establishes new state-of-the-art in Basque and Swedish in the SPMRL shared tasks on multilingual constituency parsing.

Efficient Second-Order TreeCRF for Neural Dependency Parsing [Website][PDF]
Yu Zhang, Zhenghua Li, and Min Zhang 5:00–6:00

In the deep learning (DL) era, parsing models are extremely simplified with little hurt on performance, thanks to the remarkable capability of multi-layer BiLSTMs in context representation. As the most popular graph-based dependency parser due to its high efficiency and performance, the biaffine parser directly scores single dependencies under the arc-factorization assumption, and adopts a very simple local token-wise cross-entropy training loss. This paper for the first time presents a second-order TreeCRF extension to the biaffine parser. For a long time, the complexity and inefficiency of the inside-outside algorithm hinder the popularity of TreeCRF. To address this issue, we propose an effective way to batchify the inside and Viterbi algorithms for direct large matrix operation on GPUs, and to avoid the complex outside algorithm via efficient back-propagation. Experiments and analysis on 27 datasets from 13 languages clearly show that techniques developed before the DL era, such as structural learning (global TreeCRF loss) and high-order modeling are still useful, and can further boost parsing performance over the state-of-the-art biaffine parser, especially for partially annotated training data. We release our code at https://github.com/yzhangcs/crfpar.

Representations of Syntax [MASK] Useful: Effects of Constituency and Dependency Structure in Recursive LSTMs [Website][PDF]
Michael Lepori, Tal Linzen, and R. Thomas McCoy 5:00–6:00

Sequence-based neural networks show significant sensitivity to syntactic structure, but they still perform less well on syntactic tasks than tree-based networks. Such tree-based networks can be provided with a constituency parse, a dependency parse, or both. We evaluate which of these two representational schemes more effectively introduces biases for syntactic structure that increase performance on the subject-verb agreement prediction task. We find that a constituency-based network generalizes more robustly than a dependency-based one, and that combining the two types of structure does not yield further improvement. Finally, we show that the syntactic robustness of sequential models can be substantially improved by fine-tuning on a small amount of constructed data, suggesting that data augmentation is a viable alternative to explicit constituency structure for imparting the syntactic biases that sequential models are lacking.

Structure-Level Knowledge Distillation For Multilingual Sequence Labeling [Website] [PDF]
Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Fei Huang, and Kewei Tu 5:00–6:00

Multilingual sequence labeling is a task of predicting label sequences using a single unified model for multiple languages. Compared with relying on multiple monolingual models, using a multilingual model has the benefit of a smaller model size, easier in online serving, and generalizability to low-resource languages. However, current multilingual models still underperform individual monolingual models significantly due to model capacity limitations. In this paper, we propose to reduce the gap between monolingual models and the unified multilingual model by distilling the structural knowledge of several monolingual models (teachers) to the unified multilingual model (student). We propose two novel KD methods based on structure-level information: (1) approximately minimizes the distance between the student's and the teachers' structure-level probability distributions, (2) aggregates the structure-level knowledge to local distributions and minimizes the distance between two local probability distributions. Our experiments on 4 multilingual tasks with 25 datasets show that our approaches outperform several strong baselines and have stronger zero-shot generalizability than both the baseline model and teacher models.

Demo Session 1B

Time: *5:45–6:30*

CLIReval: Evaluating Machine Translation as a Cross-Lingual Information Retrieval Task [Website][PDF]

Shuo Sun, Suzanna Sia, and Kevin Duh

We present CLIReval, an easy-to-use toolkit for evaluating machine translation (MT) with the proxy task of cross-lingual information retrieval (CLIR). Contrary to what the project name might suggest, CLIReval does not actually require any annotated CLIR dataset. Instead, it automatically transforms translations and references used in MT evaluations into a synthetic CLIR dataset; it then sets up a standard search engine (Elasticsearch) and computes various information retrieval metrics (e.g., mean average precision) by treating the translations as documents to be retrieved. The idea is to gauge the quality of MT by its impact on the document translation approach to CLIR. As a case study, we run CLIReval on the "metrics shared task" of WMT2019; while this extrinsic metric is not intended to replace popular intrinsic metrics such as BLEU, results suggest CLIReval is competitive in many language pairs in terms of correlation to human judgments of quality. CLIReval is publicly available at https://github.com/ssun32/CLIReval.

Session 6B Overview – Tuesday, July 7, 2020 6:00–7:00

Track A *Computational Social Science and Social Media-4* Abstracts	Dynamic Online Conversation Recommendation *Zeng, Li, Wang, Mao, and Wong* [Website][PDF]	Improving Multimodal Named Entity Recognition via Entity Span Detection with Unified Multimodal Transformer *Yu, Jiang, Yang, and Xia* [Website][PDF]	Stock Embeddings Acquired from News Articles and Price History, and an Application to Portfolio Optimization *Du and Tanaka-Ishii* [Website][PDF]	What Was Written vs. Who Read It: News Media Profiling Using Text Analysis and Social Media Context *Baly, Karadzhov, An, Kwak, Dinkov, Ali, Glass, and Nakov* [Website][PDF]	
Track B *Interpretability and Analysis of Models for NLP-1* Abstracts	An Analysis of the Utility of Explicit Negative Examples to Improve the Syntactic Abilities of Neural Language Models *Noji and Takamura* [Website][PDF]	[TACL] Membership Inference Attacks on Sequence-to-Sequence Models: Is My Data In Your Machine Translation System? *Hisamoto, Post, and Duh* [Website][PDF]	On the Robustness of Language Encoders against Grammatical Errors *Yin, Long, Meng, and Chang* [Website][PDF]	Roles and Utilization of Attention Heads in Transformer-based Neural Language Models *Jo and Myaeng* [Website][PDF]	[TACL] Tabula nearly rasa: Probing the linguistic knowledge of character-level neural language models trained on unsegmented text *Hahn and Baroni* [Website][PDF]
	Understanding Attention for Text Classification *Sun and Lu* [Website][PDF]				
Track C *Machine Learning for NLP-6* Abstracts	A Relational Memory-based Embedding Model for Triple Classification and Search Personalization *Nguyen, Nguyen, and Phung* [Website][PDF]	Do you have the right scissors? Tailoring Pre-trained Language Models via Monte-Carlo Methods *Miao, Song, Zhou, and Li* [Website][PDF]	Enhancing Pre-trained Chinese Character Representation with Word-aligned Attention *Li, Yu, Mengge, and Liu* [Website][PDF]	On the Encoder-Decoder Incompatibility in Variational Text Modeling and Beyond *Wu, Wang, and Wang* [Website][PDF]	SAFER: A Structure-free Approach for Certified Robustness to Adversarial Word Substitutions *Ye, Gong, and Liu* [Website][PDF]
Track D *Machine Translation-8* Abstracts	A Graph-based Coarse-to-fine Method for Unsupervised Bilingual Lexicon Induction *Ren, Liu, Zhou, and Ma* [Website][PDF]	A Reinforced Generation of Adversarial Examples for Neural Machine Translation *, Huang, Xie, Dai, and CHEN* [Website][PDF]	A Retrieve-and-Rewrite Initialization Method for Unsupervised Machine Translation *Ren, Wu, Liu, Zhou, and Ma* [Website][PDF]	A Simple and Effective Unified Encoder for Document-Level Machine Translation *Ma, Zhang, and Zhou* [Website][PDF]	Does Multi-Encoder Help? A Case Study on Context-Aware Neural Machine Translation *Li, Liu, Wang, Jiang, Xiao, Zhu, Liu, and* [Website][PDF]
	Dynamically Adjusting Transformer Batch Size by Monitoring Gradient Direction Change *Xu, Genabith, Xiong, and Liu* [Website][PDF]	Knowledge Distillation for Multilingual Unsupervised Neural Machine Translation *Sun, Wang, Chen, Utiyama, Sumita, and Zhao* [Website][PDF]	Lexically Constrained Neural Machine Translation with Levenshtein Transformer *Susanto, Chollampatt, and Tan* [Website][PDF]	On Exposure Bias, Hallucination and Domain Shift in Neural Machine Translation *Wang and Sennrich* [Website][PDF]	

Track E *Resources and Evaluation-5* Abstracts	Automatic Machine Translation Evaluation using Source Language Inputs and Cross-lingual Language Model *Takahashi, Sudoh, and Nakamura* [Website][PDF]	ChartDialogs: Plotting from Natural Language Instructions *Shao and Nakashole* [Website][PDF]	GLUECoS: An Evaluation Benchmark for Code-Switched NLP *Khanuja, Dandapat, Srinivasan, Sitaram, and Choudhury* [Website][PDF]	MATINF: A Jointly Labeled Large-Scale Dataset for Classification, Question Answering and Summarization *Xu, Pei, Wu, Liu, and Li* [Website][PDF]	MIND: A Large-scale Dataset for News Recommendation *Wu, Qiao, Chen, Wu, Qi, Lian, Liu, Xie, Gao, Wu, and Zhou* [Website][PDF]
	That is a Known Lie: Detecting Previously Fact-Checked Claims *Shaar, Babulkov, Da San Martino, and Nakov* [Website][PDF]	Towards Holistic and Automatic Evaluation of Open-Domain Dialogue Generation *Pang, Nijkamp, Han, Zhou, Liu, and Tu* [Website][PDF]			
Track F *Lexical-4* Abstracts	BiRRE: Learning Bidirectional Residual Relation Embeddings for Supervised Hypernymy Detection *Wang and HE* [Website][PDF]	Biomedical Entity Representations with Synonym Marginalization *Sung, Jeon, Lee, and Kang* [Website][PDF]	Hypernymy Detection for Low-Resource Languages via Meta Learning *Yu, Han, Zhang, and Ng* [Website][PDF]	Investigating Word-Class Distributions in Word Vector Spaces *Sasano and Korhonen* [Website][PDF]	
Track G *Sentiment Analysis, Stylistic Analysis, and Argument Mining-2* Abstracts	Aspect Sentiment Classification with Document-level Sentiment Preference Modeling *Chen, Sun, Wang, Li, Si, Zhang, and Zhou* [Website][PDF]	Don't Eclipse Your Arts Due to Small Discrepancies: Boundary Repositioning with a Pointer Network for Aspect Extraction *Wei, Hong, Zou, Cheng, and YAO* [Website][PDF]	Relation-Aware Collaborative Learning for Unified Aspect-Based Sentiment Analysis *Chen and Qian* [Website][PDF]	SentiBERT: A Transferable Transformer-Based Architecture for Compositional Sentiment Semantics *Yin, Meng, and Chang* [Website][PDF]	Transition-based Directed Graph Construction for Emotion-Cause Pair Extraction *Fan, Yuan, Du, Gui, Yang, and Xu* [Website][PDF]
Track H *Speech and Multimodality-3* Abstracts	CH-SIMS: A Chinese Multimodal Sentiment Analysis Dataset with Fine-grained Annotation of Modality *Yu, Xu, Meng, Zhu, Ma, Wu, Zou, and Yang* [Website][PDF]	Curriculum Pre-training for End-to-End Speech Translation *Wang, Wu, Liu, Zhou, and Yang* [Website][PDF]	How Accents Confound: Probing for Accent Information in End-to-End Speech Recognition Systems *Prasad and Jyothi* [Website][PDF]	Improving Disfluency Detection by Self-Training a Self-Attentive Model *Jamshid Lou and Johnson* [Website][PDF]	Learning Spoken Language Representations with Neural Lattice Language Modeling *Huang and Chen* [Website][PDF]
	Meta-Transfer Learning for Code-Switched Speech Recognition *Winata, Cahyawijaya, Lin, Liu, Xu, and Fung* [Website][PDF]	Reasoning with Multimodal Sarcastic Tweets via Modeling Cross-Modality Contrast and Semantic Association *Xu, Zeng, and Mao* [Website][PDF]	SimulSpeech: End-to-End Simultaneous Speech to Text Translation *Ren, Liu, Tan, Zhang, QIN, Zhao, and Liu* [Website][PDF]	Towards end-2-end learning for predicting behavior codes from spoken utterances in psychotherapy conversations *Singla, Chen, Atkins, and Narayanan* [Website][PDF]	

Track I *Student Research Workshop* Abstracts	Embeddings of Label Components for Sequence Labeling: A Case Study of Fine-grained Named Entity Recognition *Kato, Abe, Ouchi, Miyawaki, Suzuki, and Inui* [Website] [PDF]	Building a Japanese Typo Dataset from Wikipedia's Revision History *Tanaka, Murawaki, Kawahara, and Kurohashi* [Website] [PDF]	Preventing Critical Scoring Errors in Short Answer Scoring with Confidence Estimation *Funayama, Sasaki, Matsubayashi, Mizumoto, Suzuki, Mita, and Inui* [Website] [PDF]	

Session 6B Details

Session 6B: Computational Social Science and Social Media-4

Dynamic Online Conversation Recommendation [Website][PDF]
Xingshan Zeng, Jing Li, Lu Wang, Zhiming Mao, and Kam-Fai Wong 6:00–7:00

Trending topics in social media content evolve over time, and it is therefore crucial to understand social media users and their interpersonal communications in a dynamic manner. Here we study dynamic online conversation recommendation, to help users engage in conversations that satisfy their evolving interests. While most prior work assumes static user interests, our model is able to capture the temporal aspects of user interests, and further handle future conversations that are unseen during training time. Concretely, we propose a neural architecture to exploit changes of user interactions and interests over time, to predict which discussions they are likely to enter. We conduct experiments on large-scale collections of Reddit conversations, and results on three subreddits show that our model significantly outperforms state-of-the-art models that make a static assumption of user interests. We further evaluate on handling "cold start", and observe consistently better performance by our model when considering various degrees of sparsity of user's chatting history and conversation contexts. Lastly, analyses on our model outputs indicate user interest change, explaining the advantage and efficacy of our approach.

Improving Multimodal Named Entity Recognition via Entity Span Detection with Unified Multimodal Transformer [Website][PDF]
Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia 6:00–7:00

In this paper, we study Multimodal Named Entity Recognition (MNER) for social media posts. Existing approaches for MNER mainly suffer from two drawbacks: (1) despite generating word-aware visual representations, their word representations are insensitive to the visual context; (2) most of them ignore the bias brought by the visual context. To tackle the first issue, we propose a multimodal interaction module to obtain both image-aware word representations and word-aware visual representations. To alleviate the visual bias, we further propose to leverage purely text-based entity span detection as an auxiliary module, and design a Unified Multimodal Transformer to guide the final predictions with the entity span predictions. Experiments show that our unified approach achieves the new state-of-the-art performance on two benchmark datasets.

Stock Embeddings Acquired from News Articles and Price History, and an Application to Portfolio Optimization [Website][PDF]
Xin Du and Kumiko Tanaka-Ishii 6:00–7:00

Previous works that integrated news articles to better process stock prices used a variety of neural networks to predict price movements. The textual and price information were both encoded in the neural network, and it is therefore difficult to apply this approach in situations other than the original framework of the notoriously hard problem of price prediction. In contrast, this paper presents a method to encode the influence of news articles through a vector representation of stocks called a *stock embedding*. The stock embedding is acquired with a deep learning framework using both news articles and price history. Because the embedding takes the operational form of a vector, it is applicable to other financial problems besides price prediction. As one example application, we show the results of portfolio optimization using Reuters & Bloomberg headlines, producing a capital gain 2.8 times larger than that obtained with a baseline method using only stock price data. This suggests that the proposed stock embedding can leverage textual financial semantics to solve financial prediction problems.

What Was Written vs. Who Read It: News Media Profiling Using Text Analysis and Social Media Context [Website][PDF]
Ramy Baly, Georgi Karadzhov, Jisun An, Haewoon Kwak, Yoan Dinkov, Ahmed Ali, James Glass, and Preslav Nakov 6:00–7:00

Predicting the political bias and the factuality of reporting of entire news outlets are critical elements of media profiling, which is an understudied but an increasingly important research direction. The present level of proliferation of fake, biased, and propagandistic content online has made it impossible to fact-check every single suspicious claim, either manually or automatically. Thus, it has been proposed to profile entire news outlets and to look for those that are likely to publish fake or biased content. This makes it possible to detect likely "fake news" the moment they are published, by simply checking the reliability of their source. From a practical perspective, political bias and factuality of reporting have a linguistic aspect but also a social context. Here, we study the impact of both, namely (i) what was written (i.e., what was published by the target medium, and how it describes itself in Twitter) vs. (ii) who reads it (i.e., analyzing the target medium's audience on social media). We further study (iii) what was written about the target medium (in Wikipedia). The evaluation results show that what was written matters most, and we further show that putting all information sources together yields huge improvements over the current state-of-the-art.

Session 6B: Interpretability and Analysis of Models for NLP-1

An Analysis of the Utility of Explicit Negative Examples to Improve the Syntactic Abilities of Neural Language Models [Website][PDF]
Hiroshi Noji and Hiroya Takamura 6:00–7:00

We explore the utilities of explicit negative examples in training neural language models. Negative examples here are incorrect words in a sentence, such as *barks* in **The dogs barks.* Neural language models are commonly trained only on positive examples, a set of sentences in the training data, but recent studies suggest that the models trained in this way are not capable of robustly handling complex syntactic constructions, such as long-distance agreement. In this paper, we first demonstrate that appropriately using negative examples about particular constructions (e.g., subject-verb agreement) will boost the model's robustness on them in English, with a negligible loss of perplexity. The key to our success is an additional margin loss between the log-likelihoods of a correct word and an incorrect word. We then provide a detailed analysis of the trained models. One of our findings is the difficulty of object-relative clauses for RNNs. We find that even with our direct learning signals the models still suffer from resolving agreement across an object-relative clause. Augmentation of training sentences involving the constructions somewhat helps, but the accuracy still does not reach the level of subject-relative clauses. Although not directly cognitively appealing, our method can be a tool to analyze the true architectural limitation of neural models on challenging linguistic constructions.

[TACL] Membership Inference Attacks on Sequence-to-Sequence Models: Is My Data In Your Machine Translation System? [Website][PDF]
Sorami Hisamoto, Matt Post, and Kevin Duh 6:00–7:00

Data privacy is an important issue for "machine learning as a service" providers. We focus on the problem of membership inference attacks: given a data sample and black-box access to a model's API, determine whether the sample existed in the model's training data. Our contribution is an investigation of this problem in the context of sequence-to-sequence models, which are important in applications such as machine translation and video captioning. We define the membership inference problem for sequence generation, provide an open dataset based on state-of-the-art machine translation models, and report initial results on whether these models leak private information against several kinds of membership inference attacks.

On the Robustness of Language Encoders against Grammatical Errors [Website][PDF]
Fan Yin, Quanyu Long, Tao Meng, and Kai-Wei Chang 6:00–7:00

We conduct a thorough study to diagnose the behaviors of pre-trained language encoders (ELMo, BERT, and RoBERTa) when confronted with natural grammatical errors. Specifically, we collect real grammatical errors from non-native speakers and conduct adversarial attacks to simulate these errors on clean text data. We use this approach to facilitate debugging models on downstream applications. Results confirm that the performance of all tested models is affected but the degree of impact varies. To interpret model behaviors, we further design a linguistic acceptability task to reveal their abilities in identifying ungrammatical sentences and the position of errors. We find that fixed contextual encoders with a simple classifier trained on the prediction of sentence correctness are able to locate error positions. We also design a cloze test for BERT and discover that BERT captures the interaction between errors and specific tokens in context. Our results shed light on understanding the robustness and behaviors of language encoders against grammatical errors.

Roles and Utilization of Attention Heads in Transformer-based Neural Language Models [Website][PDF]
Jae-young Jo and Sung-Hyon Myaeng 6:00–7:00

Sentence encoders based on the transformer architecture have shown promising results on various natural language tasks. The main impetus lies in the pre-trained neural language models that capture long-range dependencies among words, owing to multi-head attention that is unique in the architecture. However, little is known for how linguistic properties are processed, represented, and utilized for downstream tasks among hundreds of attention heads inside the pre-trained transformer-based model. For the initial goal of examining the roles of attention heads in handling a set of linguistic features, we conducted a set of experiments with ten probing tasks and three downstream tasks on four pre-trained transformer families (GPT, GPT2, BERT, and ELECTRA). Meaningful insights are shown through the lens of heat map visualization and utilized to propose a relatively simple sentence representation method that takes advantage of most influential attention heads, resulting in additional performance improvements on the downstream tasks.

[TACL] Tabula nearly rasa: Probing the linguistic knowledge of character-level neural language models trained on unsegmented text [Website][PDF]
Michael Hahn and Marco Baroni 6:00–7:00

Recurrent neural networks (RNNs) reached striking performance in many natural language processing tasks. This has renewed interest in whether these generic sequence processing devices are inducing genuine linguistic knowledge. Nearly all current analytical studies, however, initialize the RNNs with a vocabulary of known words, and feed them tokenized input during training. We present a multi-lingual study of the linguistic knowledge encoded in RNNs trained as character-level language models, on input data with word boundaries removed. These networks face a tougher and more cognitively realistic task, having to discover and store any useful linguistic unit from scratch, based on input statistics. The results show that our "near tabula rasa" RNNs are mostly able to solve morphological, syntactic and semantic tasks that intuitively presuppose word-level knowledge, and indeed they learned to track "soft" word boundaries. Our study opens the door to speculations about the necessity of an explicit word lexicon in language learning and usage.

Understanding Attention for Text Classification [Website][PDF]
Xiaobing Sun and Wei Lu 6:00–7:00

Attention has been proven successful in many natural language processing (NLP) tasks. Recently, many researchers started to investigate the interpretability of attention on NLP tasks. Many existing approaches focused on examining whether the local attention weights could reflect the importance of input representations. In this work, we present a study on understanding the internal mechanism of attention by looking into the gradient update process, checking its behavior when approaching a local minimum during training. We propose to analyze for each word token the following two quantities: its polarity score and its attention score, where the latter is a global assessment on the token's significance. We discuss conditions under which the attention mechanism may become more (or less) interpretable, and show how the interplay between the two quantities can contribute towards model performance.

Session 6B: Machine Learning for NLP-6

A Relational Memory-based Embedding Model for Triple Classification and Search Personalization [Website][PDF]
Dai Quoc Nguyen, Tu Nguyen, and Dinh Phung 6:00–7:00

Knowledge graph embedding methods often suffer from a limitation of memorizing valid triples to predict new ones for triple classification and search personalization problems. To this end, we introduce a novel embedding model, named R-MeN, that explores a relational memory network to encode potential dependencies in relationship triples. R-MeN considers each triple as a sequence of 3 input vectors that recurrently interact with a memory using a transformer self-attention mechanism. Thus R-MeN encodes new information from interactions between the memory and each input vector to return a corresponding vector. Consequently, R-MeN feeds these 3 returned vectors to a convolutional neural network-based decoder to produce a scalar score for the triple. Experimental results show that our proposed R-MeN obtains state-of-the-art results on SEARCH17 for the search personalization task, and on WN11 and FB13 for the triple classification task.

Do you have the right scissors? Tailoring Pre-trained Language Models via Monte-Carlo Methods [Website][PDF]
Ning Miao, Yuxuan Song, Hao Zhou, and Lei Li 6:00–7:00

It has been a common approach to pre-train a language model on a large corpus and fine-tune it on task-specific data. In practice, we observe that fine-tuning a pre-trained model on a small dataset may lead to over- and/or under-estimate problem. In this paper, we propose MC-Tailor, a novel method to alleviate the above issue in text generation tasks by truncating and transferring the probability mass from over-estimated regions to under-estimated ones. Experiments on a variety of text generation datasets show that MC-Tailor consistently and significantly outperforms the fine-tuning approach.

Enhancing Pre-trained Chinese Character Representation with Word-aligned Attention [Website][PDF]
Yanzeng Li, Bowen Yu, Xue Mengge, and Tingwen Liu 6:00–7:00

Most Chinese pre-trained models take character as the basic unit and learn representation according to character's external contexts, ignoring the semantics expressed in the word, which is the smallest meaningful utterance in Chinese. Hence, we propose a novel word-aligned attention to exploit explicit word information, which is complementary to various character-based Chinese pre-trained language models. Specifically, we devise a pooling mechanism to align the character-level attention to the word level and propose to alleviate the potential issue of segmentation error propagation by multi-source information fusion. As a result, word and character information are explicitly integrated at the fine-tuning procedure. Experimental results on five Chinese NLP benchmark tasks demonstrate that our method achieves significant improvements against BERT, ERNIE and BERT-wwm.

On the Encoder-Decoder Incompatibility in Variational Text Modeling and Beyond [Website][PDF]
Chen Wu, Prince Zizhuang Wang, and William Yang Wang 6:00–7:00

Variational autoencoders (VAEs) combine latent variables with amortized variational inference, whose optimization usually converges into a trivial local optimum termed posterior collapse, especially in text modeling. By tracking the optimization dynamics, we observe the encoder-decoder incompatibility that leads to poor parameterizations of the data manifold. We argue that the trivial local optimum may be avoided by improving the encoder and decoder parameterizations since the posterior network is part of a transition map between them. To this end, we propose Coupled-VAE, which couples a VAE model with a deterministic autoencoder with the same structure and improves the encoder and decoder parameterizations via encoder weight sharing and decoder signal matching. We apply the proposed Coupled-VAE approach to various VAE models with different regularization, posterior family, decoder structure, and optimization strategy. Experiments on benchmark datasets (i.e., PTB, Yelp, and Yahoo) show consistently improved results in terms of probability estimation and richness of the latent space. We also generalize our method to conditional language modeling and propose Coupled-CVAE, which largely improves the diversity of dialogue generation on the Switchboard dataset.

SAFER: A Structure-free Approach for Certified Robustness to Adversarial Word Substitutions [Website][PDF]
Mao Ye, Chengyue Gong, and Qiang Liu 6:00–7:00

State-of-the-art NLP models can often be fooled by human-unaware transformations such as synonymous word substitution. For security reasons, it is of critical importance to develop models with certified robustness that can provably guarantee that the prediction is can not be altered by any possible synonymous word substitution. In this work, we propose a certified robust method based on a new randomized smoothing technique, which constructs a stochastic ensemble by applying random word substitutions on the input sentences, and leverage the statistical properties of the ensemble to provably certify the robustness. Our method is simple and structure-free in that it only requires the black-box queries of the model outputs, and hence can be applied to any pre-trained models (such as BERT) and any types of models (world-level or subword-level). Our method significantly outperforms recent state-of-the-art methods for certified robustness on both IMDB and Amazon text classification tasks. To the best of our knowledge, we are the first work to achieve certified robustness on large systems such as BERT with practically meaningful certified accuracy.

Session 6B: Machine Translation-8

A Graph-based Coarse-to-fine Method for Unsupervised Bilingual Lexicon Induction [Website][PDF]
Shuo Ren, Shujie Liu, Ming Zhou, and Shuai Ma 6:00–7:00

Unsupervised bilingual lexicon induction is the task of inducing word translations from monolingual corpora of two languages. Recent methods are mostly based on unsupervised cross-lingual word embeddings, the key to which is to find initial solutions of word translations, followed by the learning and refinement of mappings between the embedding spaces of two languages. However, previous methods find initial solutions just based on word-level information, which may be (1) limited and inaccurate, and (2) prone to contain some noise introduced by the insufficiently pre-trained embeddings of some words. To deal with those issues, in this paper, we propose a novel graph-based paradigm to induce bilingual lexicons in a coarse-to-fine way. We first build a graph for each language with its vertices representing different words. Then we extract word cliques from the graphs and map the cliques of two languages. Based on that, we induce the initial word translation solution with the central words of the aligned cliques. This coarse-to-fine approach not only leverages clique-level information, which is richer and more accurate, but also effectively reduces the bad effect of the noise in the pre-trained embeddings. Finally, we take the initial solution as the seed to learn cross-lingual embeddings, from which we induce bilingual lexicons. Experiments show that our approach improves the performance of bilingual lexicon induction compared with previous methods.

A Reinforced Generation of Adversarial Examples for Neural Machine Translation [Website][PDF]
wei zou wei, Shujian Huang, Jun Xie, Xinyu Dai, and Jiajun CHEN 6:00–7:00

Neural machine translation systems tend to fail on less decent inputs despite its significant efficacy, which may significantly harm the credibility of these systems—fathoming how and when neural-based systems fail in such cases is critical for industrial maintenance. Instead of collecting and analyzing bad cases using limited handcrafted error features, here we investigate this issue by generating adversarial examples via a new paradigm based on reinforcement learning. Our paradigm could expose pitfalls for a given performance metric, e.g., BLEU, and could target any given neural machine translation architecture. We conduct experiments of adversarial attacks on two mainstream neural machine translation architectures, RNN-search, and Transformer. The results show that our method efficiently produces stable attacks with meaning-preserving adversarial examples. We also present a qualitative and quantitative analysis for the preference pattern of the attack, demonstrating its capability of pitfall exposure.

A Retrieve-and-Rewrite Initialization Method for Unsupervised Machine Translation [Website][PDF]
Shuo Ren, Yu Wu, Shujie Liu, Ming Zhou, and Shuai Ma 6:00–7:00

The commonly used framework for unsupervised machine translation builds initial translation models of both translation directions, and then performs iterative back-translation to jointly boost their translation performance. The initialization stage is very important since bad initialization may wrongly squeeze the search space, and too much noise introduced in this stage may hurt the final performance. In this paper, we propose a novel retrieval and rewriting based method to better initialize unsupervised translation models. We first retrieve semantically comparable sentences from monolingual corpora of two languages and then rewrite the target side to minimize the semantic gap between the source and retrieved targets with a designed rewriting model. The rewritten sentence pairs are used to initialize SMT models which are used to generate pseudo data for two NMT models, followed by the iterative back-translation. Experiments show that our method can build better initial unsupervised translation models and improve the final translation performance by over 4 BLEU scores. Our code is released at https://github.com/Imagist-Shuo/RRforUNMT.git.

A Simple and Effective Unified Encoder for Document-Level Machine Translation [Website][PDF]
Shuming Ma, Dongdong Zhang, and Ming Zhou 6:00–7:00

Most of the existing models for document-level machine translation adopt dual-encoder structures. The representation of the source sentences and the document-level contexts[1] are modeled with two separate encoders. Although these models can make use of the document-level contexts, they do not fully model the interaction between the contexts and the source sentences, and can not directly adapt to the recent pre-training models (e.g., BERT) which encodes multiple sentences with a single encoder. In this work, we propose a simple and effective unified encoder that can outperform the baseline models of dual-encoder models in terms of BLEU and METEOR scores. Moreover, the pre-training models can further boost the performance of our proposed model.

Does Multi-Encoder Help? A Case Study on Context-Aware Neural Machine Translation [Website][PDF]
Bei Li, Hui Liu, Ziyang Wang, Yufan Jiang, Tong Xiao, Jingbo Zhu, Tongran Liu, and changliang li changliang 6:00–7:00

In encoder-decoder neural models, multiple encoders are in general used to represent the contextual information in addition to the individual sentence. In this paper, we investigate multi-encoder approaches in document-level neural machine translation (NMT). Surprisingly, we find that the context encoder does not only encode the surrounding sentences but also behaves as a noise generator. This makes us rethink the real benefits of multi-encoder in context-aware translation - some of the improvements come from robust training. We compare several methods that introduce noise and/or well-tuned dropout setup into the training of these encoders. Experimental results show that noisy training plays an important role in multi-encoder-based NMT, especially when the training data is small. Also, we establish a new state-of-the-art on IWSLT Fr-En task by careful use of noise generation and dropout methods.

[1] In this work, document-level contexts denote the surrounding sentences of the current source sentence.

Dynamically Adjusting Transformer Batch Size by Monitoring Gradient Direction Change [Website][PDF]
Hongfei Xu, Josef van Genabith, Deyi Xiong, and Qiuhui Liu 6:00–7:00

The choice of hyper-parameters affects the performance of neural models. While much previous research (Sutskever et al., 2013; Duchi et al., 2011; Kingma and Ba, 2015) focuses on accelerating convergence and reducing the effects of the learning rate, comparatively few papers concentrate on the effect of batch size. In this paper, we analyze how increasing batch size affects gradient direction, and propose to evaluate the stability of gradients with their angle change. Based on our observations, the angle change of gradient direction first tends to stabilize (i.e. gradually decrease) while accumulating mini-batches, and then starts to fluctuate. We propose to automatically and dynamically determine batch sizes by accumulating gradients of mini-batches and performing an optimization step at just the time when the direction of gradients starts to fluctuate. To improve the efficiency of our approach for large models, we propose a sampling approach to select gradients of parameters sensitive to the batch size. Our approach dynamically determines proper and efficient batch sizes during training. In our experiments on the WMT 14 English to German and English to French tasks, our approach improves the Transformer with a fixed 25k batch size by +0.73 and +0.82 BLEU respectively.

Knowledge Distillation for Multilingual Unsupervised Neural Machine Translation [Website][PDF]
Haipeng Sun, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao 6:00–7:00

Unsupervised neural machine translation (UNMT) has recently achieved remarkable results for several language pairs. However, it can only translate between a single language pair and cannot produce translation results for multiple language pairs at the same time. That is, research on multilingual UNMT has been limited. In this paper, we empirically introduce a simple method to translate between thirteen languages using a single encoder and a single decoder, making use of multilingual data to improve UNMT for all language pairs. On the basis of the empirical findings, we propose two knowledge distillation methods to further enhance multilingual UNMT performance. Our experiments on a dataset with English translated to and from twelve other languages (including three language families and six language branches) show remarkable results, surpassing strong unsupervised individual baselines while achieving promising performance between non-English language pairs in zero-shot translation scenarios and alleviating poor performance in low-resource language pairs.

Lexically Constrained Neural Machine Translation with Levenshtein Transformer [Website][PDF]
Raymond Hendy Susanto, Shamil Chollampatt, and Liling Tan 6:00–7:00

This paper proposes a simple and effective algorithm for incorporating lexical constraints in neural machine translation. Previous work either required re-training existing models with the lexical constraints or incorporating them during beam search decoding with significantly higher computational overheads. Leveraging the flexibility and speed of a recently proposed Levenshtein Transformer model (Gu et al., 2019), our method injects terminology constraints at inference time without any impact on decoding speed. Our method does not require any modification to the training procedure and can be easily applied at runtime with custom dictionaries. Experiments on English-German WMT datasets show that our approach improves an unconstrained baseline and previous approaches.

On Exposure Bias, Hallucination and Domain Shift in Neural Machine Translation [Website][PDF]
Chaojun Wang and Rico Sennrich 6:00–7:00

The standard training algorithm in neural machine translation (NMT) suffers from exposure bias, and alternative algorithms have been proposed to mitigate this. However, the practical impact of exposure bias is under debate. In this paper, we link exposure bias to another well-known problem in NMT, namely the tendency to generate hallucinations under domain shift. In experiments on three datasets with multiple test domains, we show that exposure bias is partially to blame for hallucinations, and that training with Minimum Risk Training, which avoids exposure bias, can mitigate this. Our analysis explains why exposure bias is more problematic under domain shift, and also links exposure bias to the beam search problem, i.e. performance deterioration with increasing beam size. Our results provide a new justification for methods that reduce exposure bias: even if they do not increase performance on in-domain test sets, they can increase model robustness to domain shift.

Session 6B: Resources and Evaluation-5

Automatic Machine Translation Evaluation using Source Language Inputs and Cross-lingual Language Model [Website][PDF]
Kosuke Takahashi, Katsuhito Sudoh, and Satoshi Nakamura 6:00–7:00

We propose an automatic evaluation method of machine translation that uses source language sentences regarded as additional pseudo references. The proposed method evaluates a translation hypothesis in a regression model. The model takes the paired source, reference, and hypothesis sentence all together as an input. A pretrained large scale cross-lingual language model encodes the input to sentence-pair vectors, and the model predicts a human evaluation score with those vectors. Our experiments show that our proposed method using Cross-lingual Language Model (XLM) trained with a translation language modeling (TLM) objective achieves a higher correlation with human judgments than a baseline method that uses only hypothesis and reference sentences. Additionally, using source sentences in our proposed method is confirmed to improve the evaluation performance.

ChartDialogs: Plotting from Natural Language Instructions [Website][PDF]
Yutong Shao and Ndapa Nakashole 6:00–7:00

This paper presents the problem of conversational plotting agents that carry out plotting actions from natural language instructions. To facilitate the development of such agents, we introduce ChartDialogs, a new multi-turn dialog dataset, covering a popular plotting library, matplotlib. The dataset contains over 15,000 dialog turns from 3,200 dialogs covering the majority of matplotlib plot types. Extensive experiments show the best-performing method achieving 61% plotting accuracy, demonstrating that the dataset presents a non-trivial challenge for future research on this task.

GLUECoS: An Evaluation Benchmark for Code-Switched NLP [Website][PDF]
Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury 6:00–7:00

Code-switching is the use of more than one language in the same conversation or utterance. Recently, multilingual contextual embedding models, trained on multiple monolingual corpora, have shown promising results on cross-lingual and multilingual tasks. We present an evaluation benchmark, GLUECoS, for code-switched languages, that spans several NLP tasks in English-Hindi and English-Spanish. Specifically, our evaluation benchmark includes Language Identification from text, POS tagging, Named Entity Recognition, Sentiment Analysis, Question Answering and a new task for code-switching, Natural Language Inference. We present results on all these tasks using cross-lingual word embedding models and multilingual models. In addition, we fine-tune multilingual models on artificially generated code-switched data. Although multilingual models perform significantly better than cross-lingual models, our results show that in most tasks, across both language pairs, multilingual models fine-tuned on code-switched data perform best, showing that multilingual models can be further optimized for code-switching tasks.

MATINF: A Jointly Labeled Large-Scale Dataset for Classification, Question Answering and Summarization [Website][PDF]
Canwen Xu, Jiaxin Pei, Hongtao Wu, Yiyu Liu, and Chenliang Li 6:00–7:00

Recently, large-scale datasets have vastly facilitated the development in nearly all domains of Natural Language Processing. However, there is currently no cross-task dataset in NLP, which hinders the development of multi-task learning. We propose MATINF, the first jointly labeled large-scale dataset for classification, question answering and summarization. MATINF contains 1.07 million question-answer pairs with human-labeled categories and user-generated question descriptions. Based on such rich information, MATINF is applicable for three major NLP tasks, including classification, question answering, and summarization. We benchmark existing methods and a novel multi-task baseline over MATINF to inspire further research. Our comprehensive comparison and experiments over MATINF and other datasets demonstrate the merits held by MATINF.

MIND: A Large-scale Dataset for News Recommendation [Website][PDF]
Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, and Ming Zhou 6:00–7:00

News recommendation is an important technique for personalized news service. Compared with product and movie recommendations which have been comprehensively studied, the research on news recommendation is much more limited, mainly due to the lack of a high-quality benchmark dataset. In this paper, we present a large-scale dataset named MIND for news recommendation. Constructed from the user click logs of Microsoft News, MIND contains 1 million users and more than 160k English news articles, each of which has rich textual content such as title, abstract and body. We demonstrate MIND a good testbed for news recommendation through a comparative study of several state-of-the-art news recommendation methods which are originally developed on different proprietary datasets. Our results show the performance of news recommendation highly relies on the quality of news content understanding and user interest modeling. Many natural language processing techniques such as effective text representation methods and pre-trained language models can effectively improve the performance of news recommendation. The MIND dataset will be available at https://msnews.github.io.

That is a Known Lie: Detecting Previously Fact-Checked Claims [Website][PDF]
Shaden Shaar, Nikolay Babulkov, Giovanni Da San Martino, and Preslav Nakov 6:00–7:00

The recent proliferation of "fake news" has triggered a number of responses, most notably the emergence of several manual fact-checking initiatives. As a result and over time, a large number of fact-checked claims have been accumu-

lated, which increases the likelihood that a new claim in social media or a new statement by a politician might have already been fact-checked by some trusted fact-checking organization, as viral claims often come back after a while in social media, and politicians like to repeat their favorite statements, true or false, over and over again. As manual fact-checking is very time-consuming (and fully automatic fact-checking has credibility issues), it is important to try to save this effort and to avoid wasting time on claims that have already been fact-checked. Interestingly, despite the importance of the task, it has been largely ignored by the research community so far. Here, we aim to bridge this gap. In particular, we formulate the task and we discuss how it relates to, but also differs from, previous work. We further create a specialized dataset, which we release to the research community. Finally, we present learning-to-rank experiments that demonstrate sizable improvements over state-of-the-art retrieval and textual similarity approaches.

Towards Holistic and Automatic Evaluation of Open-Domain Dialogue Generation [Website] [PDF]
Bo Pang, Erik Nijkamp, Wenjuan Han, Linqi Zhou, Yixian Liu, and Kewei Tu 6:00–7:00

Open-domain dialogue generation has gained increasing attention in Natural Language Processing. Its evaluation requires a holistic means. Human ratings are deemed as the gold standard. As human evaluation is inefficient and costly, an automated substitute is highly desirable. In this paper, we propose holistic evaluation metrics that capture different aspects of open-domain dialogues. Our metrics consist of (1) GPT-2 based context coherence between sentences in a dialogue, (2) GPT-2 based fluency in phrasing, (3) n-gram based diversity in responses to augmented queries, and (4) textual-entailment-inference based logical self-consistency. The empirical validity of our metrics is demonstrated by strong correlations with human judgments. We open source the code and relevant materials.

Session 6B Semantics: Lexical-4

BiRRE: Learning Bidirectional Residual Relation Embeddings for Supervised Hypernymy Detection
[Website][PDF]
Chengyu Wang and XIAOFENG HE 6:00–7:00

The hypernymy detection task has been addressed under various frameworks. Previously, the design of unsupervised hypernymy scores has been extensively studied. In contrast, supervised classifiers, especially distributional models, leverage the global contexts of terms to make predictions, but are more likely to suffer from "lexical memorization". In this work, we revisit supervised distributional models for hypernymy detection. Rather than taking embeddings of two terms as classification inputs, we introduce a representation learning framework named Bidirectional Residual Relation Embeddings (BiRRE). In this model, a term pair is represented by a BiRRE vector as features for hypernymy classification, which models the possibility of a term being mapped to another in the embedding space by hypernymy relations. A Latent Projection Model with Negative Regularization (LPMNR) is proposed to simulate how hypernyms and hyponyms are generated by neural language models, and to generate BiRRE vectors based on bidirectional residuals of projections. Experiments verify BiRRE outperforms strong baselines over various evaluation frameworks.

Biomedical Entity Representations with Synonym Marginalization [Website][PDF]
Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang 6:00–7:00

Biomedical named entities often play important roles in many biomedical text mining tools. However, due to the incompleteness of provided synonyms and numerous variations in their surface forms, normalization of biomedical entities is very challenging. In this paper, we focus on learning representations of biomedical entities solely based on the synonyms of entities. To learn from the incomplete synonyms, we use a model-based candidate selection and maximize the marginal likelihood of the synonyms present in top candidates. Our model-based candidates are iteratively updated to contain more difficult negative samples as our model evolves. In this way, we avoid the explicit pre-selection of negative samples from more than 400K candidates. On four biomedical entity normalization datasets having three different entity types (disease, chemical, adverse reaction), our model BioSyn consistently outperforms previous state-of-the-art models almost reaching the upper bound on each dataset.

Hypernymy Detection for Low-Resource Languages via Meta Learning [Website][PDF]
Changlong Yu, Jialong Han, Haisong Zhang, and Wilfred Ng 6:00–7:00

Hypernymy detection, a.k.a, lexical entailment, is a fundamental sub-task of many natural language understanding tasks. Previous explorations mostly focus on monolingual hypernymy detection on high-resource languages, e.g., English, but few investigate the low-resource scenarios. This paper addresses the problem of low-resource hypernymy detection by combining high-resource languages. We extensively compare three joint training paradigms and for the first time propose applying meta learning to relieve the low-resource issue. Experiments demonstrate the superiority of our method among the three settings, which substantially improves the performance of extremely low-resource languages by preventing over-fitting on small datasets.

Investigating Word-Class Distributions in Word Vector Spaces [Website][PDF]
Ryohei Sasano and Anna Korhonen 6:00–7:00

This paper presents an investigation on the distribution of word vectors belonging to a certain word class in a pre-trained word vector space. To this end, we made several assumptions about the distribution, modeled the distribution accordingly, and validated each assumption by comparing the goodness of each model. Specifically, we considered two types of word classes — the semantic class of direct objects of a verb and the semantic class in a thesaurus — and tried to build models that properly estimate how likely it is that a word in the vector space is a member of a given word class. Our results on selectional preference and WordNet datasets show that the centroid-based model will fail to achieve good enough performance, the geometry of the distribution and the existence of subgroups will have limited impact, and also the negative instances need to be considered for adequate modeling of the distribution. We further investigated the relationship between the scores calculated by each model and the degree of membership and found that discriminative learning-based models are best in finding the boundaries of a class, while models based on the offset between positive and negative instances perform best in determining the degree of membership.

Session 6B: Sentiment Analysis, Stylistic Analysis, and Argument Mining-2

Aspect Sentiment Classification with Document-level Sentiment Preference Modeling [Website] [PDF]
Xiao Chen, Changlong Sun, Jingjing Wang, Shoushan Li, Luo Si, Min Zhang, and Guodong Zhou 6:00–7:00

In the literature, existing studies always consider Aspect Sentiment Classification (ASC) as an independent sentence-level classification problem aspect by aspect, which largely ignore the document-level sentiment preference information, though obviously such information is crucial for alleviating the information deficiency problem in ASC. In this paper, we explore two kinds of sentiment preference information inside a document, i.e., contextual sentiment consistency w.r.t. the same aspect (namely intra-aspect sentiment consistency) and contextual sentiment tendency w.r.t. all the related aspects (namely inter-aspect sentiment tendency). On the basis, we propose a Cooperative Graph Attention Networks (CoGAN) approach for cooperatively learning the aspect-related sentence representation. Specifically, two graph attention networks are leveraged to model above two kinds of document-level sentiment preference information respectively, followed by an interactive mechanism to integrate the two-fold preference. Detailed evaluation demonstrates the great advantage of the proposed approach to ASC over the state-of-the-art baselines. This justifies the importance of the document-level sentiment preference information to ASC and the effectiveness of our approach capturing such information.

Don't Eclipse Your Arts Due to Small Discrepancies: Boundary Repositioning with a Pointer Network for Aspect Extraction [Website] [PDF]
Zhenkai Wei, Yu Hong, Bowei Zou, Meng Cheng, and Jianmin YAO 6:00–7:00

The current aspect extraction methods suffer from boundary errors. In general, these errors lead to a relatively minor difference between the extracted aspects and the ground-truth. However, they hurt the performance severely. In this paper, we propose to utilize a pointer network for repositioning the boundaries. Recycling mechanism is used, which enables the training data to be collected without manual intervention. We conduct the experiments on the benchmark datasets SE14 of laptop and SE14-16 of restaurant. Experimental results show that our method achieves substantial improvements over the baseline, and outperforms state-of-the-art methods.

Relation-Aware Collaborative Learning for Unified Aspect-Based Sentiment Analysis [Website] [PDF]
Zhuang Chen and Tieyun Qian 6:00–7:00

Aspect-based sentiment analysis (ABSA) involves three subtasks, i.e., aspect term extraction, opinion term extraction, and aspect-level sentiment classification. Most existing studies focused on one of these subtasks only. Several recent researches made successful attempts to solve the complete ABSA problem with a unified framework. However, the interactive relations among three subtasks are still under-exploited. We argue that such relations encode collaborative signals between different subtasks. For example, when the opinion term is *"delicious"*, the aspect term must be *"food"* rather than *"place"*. In order to fully exploit these relations, we propose a Relation-Aware Collaborative Learning (RACL) framework which allows the subtasks to work coordinately via the multi-task learning and relation propagation mechanisms in a stacked multi-layer network. Extensive experiments on three real-world datasets demonstrate that RACL significantly outperforms the state-of-the-art methods for the complete ABSA task.

SentiBERT: A Transferable Transformer-Based Architecture for Compositional Sentiment Semantics [Website] [PDF]
Da Yin, Tao Meng, and Kai-Wei Chang 6:00–7:00

We propose SentiBERT, a variant of BERT that effectively captures compositional sentiment semantics. The model incorporates contextualized representation with binary constituency parse tree to capture semantic composition. Comprehensive experiments demonstrate that SentiBERT achieves competitive performance on phrase-level sentiment classification. We further demonstrate that the sentiment composition learned from the phrase-level annotations on SST can be transferred to other sentiment analysis tasks as well as related tasks, such as emotion classification tasks. Moreover, we conduct ablation studies and design visualization methods to understand SentiBERT. We show that SentiBERT is better than baseline approaches in capturing negation and the contrastive relation and model the compositional sentiment semantics.

Transition-based Directed Graph Construction for Emotion-Cause Pair Extraction [Website] [PDF]
Chuang Fan, Chaofa Yuan, Jiachen Du, Lin Gui, Min Yang, and Ruifeng Xu 6:00–7:00

Emotion-cause pair extraction aims to extract all potential pairs of emotions and corresponding causes from unannotated emotion text. Most existing methods are pipelined framework, which identifies emotions and extracts causes separately, leading to a drawback of error propagation. Towards this issue, we propose a transition-based model to transform the task into a procedure of parsing-like directed graph construction. The proposed model incrementally generates the directed graph with labeled edges based on a sequence of actions, from which we can recognize emotions with the corresponding causes simultaneously, thereby optimizing separate subtasks jointly and maximizing mutual benefits of tasks interdependently. Experimental results show that our approach achieves the best performance, outperforming the state-of-the-art methods by 6.71% ($p<0.01$) in F1 measure.

Session 6B: Speech and Multimodality-3

CH-SIMS: A Chinese Multimodal Sentiment Analysis Dataset with Fine-grained Annotation of Modality [Website] [PDF]
Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang 6:00–7:00

Previous studies in multimodal sentiment analysis have used limited datasets, which only contain unified multimodal annotations. However, the unified annotations do not always reflect the independent sentiment of single modalities and limit the model to capture the difference between modalities. In this paper, we introduce a Chinese single- and multi-modal sentiment analysis dataset, CH-SIMS, which contains 2,281 refined video segments in the wild with both multimodal and independent unimodal annotations. It allows researchers to study the interaction between modalities or use independent unimodal annotations for unimodal sentiment analysis.Furthermore, we propose a multi-task learning framework based on late fusion as the baseline. Extensive experiments on the CH-SIMS show that our methods achieve state-of-the-art performance and learn more distinctive unimodal representations. The full dataset and codes are available for use at https://github.com/thuiar/MMSA.

Curriculum Pre-training for End-to-End Speech Translation [Website] [PDF]
Chengyi Wang, Yu Wu, Shujie Liu, Ming Zhou, and Zhenglu Yang 6:00–7:00

End-to-end speech translation poses a heavy burden on the encoder because it has to transcribe, understand, and learn cross-lingual semantics simultaneously. To obtain a powerful encoder, traditional methods pre-train it on ASR data to capture speech features. However, we argue that pre-training the encoder only through simple speech recognition is not enough, and high-level linguistic knowledge should be considered. Inspired by this, we propose a curriculum pre-training method that includes an elementary course for transcription learning and two advanced courses for understanding the utterance and mapping words in two languages. The difficulty of these courses is gradually increasing. Experiments show that our curriculum pre-training method leads to significant improvements on En-De and En-Fr speech translation benchmarks.

How Accents Confound: Probing for Accent Information in End-to-End Speech Recognition Systems [Website] [PDF]
Archiki Prasad and Preethi Jyothi 6:00–7:00

In this work, we present a detailed analysis of how accent information is reflected in the internal representation of speech in an end-to-end automatic speech recognition (ASR) system. We use a state-of-the-art end-to-end ASR system, comprising convolutional and recurrent layers, that is trained on a large amount of US-accented English speech and evaluate the model on speech samples from seven different English accents. We examine the effects of accent on the internal representation using three main probing techniques: a) Gradient-based explanation methods, b) Information-theoretic measures, and c) Outputs of accent and phone classifiers. We find different accents exhibiting similar trends irrespective of the probing technique used. We also find that most accent information is encoded within the first recurrent layer, which is suggestive of how one could adapt such an end-to-end model to learn representations that are invariant to accents.

Improving Disfluency Detection by Self-Training a Self-Attentive Model [Website] [PDF]
Paria Jamshid Lou and Mark Johnson 6:00–7:00

Self-attentive neural syntactic parsers using contextualized word embeddings (e.g. ELMo or BERT) currently produce state-of-the-art results in joint parsing and disfluency detection in speech transcripts. Since the contextualized word embeddings are pre-trained on a large amount of unlabeled data, using additional unlabeled data to train a neural model might seem redundant. However, we show that self-training — a semi-supervised technique for incorporating unlabeled data — sets a new state-of-the-art for the self-attentive parser on disfluency detection, demonstrating that self-training provides benefits orthogonal to the pre-trained contextualized word representations. We also show that ensembling self-trained parsers provides further gains for disfluency detection.

Learning Spoken Language Representations with Neural Lattice Language Modeling [Website] [PDF]
Chao-Wei Huang and Yun-Nung Chen 6:00–7:00

Pre-trained language models have achieved huge improvement on many NLP tasks. However, these methods are usually designed for written text, so they do not consider the properties of spoken language. Therefore, this paper aims at generalizing the idea of language model pre-training to lattices generated by recognition systems. We propose a framework that trains neural lattice language models to provide contextualized representations for spoken language understanding tasks. The proposed two-stage pre-training approach reduces the demands of speech data and has better efficiency. Experiments on intent detection and dialogue act recognition datasets demonstrate that our proposed method consistently outperforms strong baselines when evaluated on spoken inputs. The code is available at https://github.com/MiuLab/Lattice-ELMo.

Meta-Transfer Learning for Code-Switched Speech Recognition [Website] [PDF]
Genta Indra Winata, Samuel Cahyawijaya, Zhaojiang Lin, Zihan Liu, Peng Xu, and Pascale Fung 6:00–7:00

An increasing number of people in the world today speak a mixed-language as a result of being multilingual. However, building a speech recognition system for code-switching remains difficult due to the availability of limited resources and the expense and significant effort required to collect mixed-language data. We therefore propose a new learning method, meta-transfer learning, to transfer learn on a code-switched speech recognition system in a low-resource setting by judiciously extracting information from high-resource monolingual datasets. Our model learns to recog-

nize individual languages, and transfer them so as to better recognize mixed-language speech by conditioning the optimization on the code-switching data. Based on experimental results, our model outperforms existing baselines on speech recognition and language modeling tasks, and is faster to converge.

Reasoning with Multimodal Sarcastic Tweets via Modeling Cross-Modality Contrast and Semantic Association [Website][PDF]
Nan Xu, Zhixiong Zeng, and Wenji Mao 6:00–7:00

Sarcasm is a sophisticated linguistic phenomenon to express the opposite of what one really means. With the rapid growth of social media, multimodal sarcastic tweets are widely posted on various social platforms. In multimodal context, sarcasm is no longer a pure linguistic phenomenon, and due to the nature of social media short text, the opposite is more often manifested via cross-modality expressions. Thus traditional text-based methods are insufficient to detect multimodal sarcasm. To reason with multimodal sarcastic tweets, in this paper, we propose a novel method for modeling cross-modality contrast in the associated context. Our method models both cross-modality contrast and semantic association by constructing the Decomposition and Relation Network (namely D&R Net). The decomposition network represents the commonality and discrepancy between image and text, and the relation network models the semantic association in cross-modality context. Experimental results on a public dataset demonstrate the effectiveness of our model in multimodal sarcasm detection.

SimulSpeech: End-to-End Simultaneous Speech to Text Translation [Website][PDF]
Yi Ren, Jinglin Liu, Xu Tan, Chen Zhang, Tao QIN, Zhou Zhao, and Tie-Yan Liu 6:00–7:00

In this work, we develop SimulSpeech, an end-to-end simultaneous speech to text translation system which translates speech in source language to text in target language concurrently. SimulSpeech consists of a speech encoder, a speech segmenter and a text decoder, where 1) the segmenter builds upon the encoder and leverages a connectionist temporal classification (CTC) loss to split the input streaming speech in real time, 2) the encoder-decoder attention adopts a wait-k strategy for simultaneous translation. SimulSpeech is more challenging than previous cascaded systems (with simultaneous automatic speech recognition (ASR) and simultaneous neural machine translation (NMT)). We introduce two novel knowledge distillation methods to ensure the performance: 1) Attention-level knowledge distillation transfers the knowledge from the multiplication of the attention matrices of simultaneous NMT and ASR models to help the training of the attention mechanism in SimulSpeech; 2) Data-level knowledge distillation transfers the knowledge from the full-sentence NMT model and also reduces the complexity of data distribution to help on the optimization of SimulSpeech. Experiments on MuST-C English-Spanish and English-German spoken language translation datasets show that SimulSpeech achieves reasonable BLEU scores and lower delay compared to full-sentence end-to-end speech to text translation (without simultaneous translation), and better performance than the two-stage cascaded simultaneous translation model in terms of BLEU scores and translation delay.

Towards end-2-end learning for predicting behavior codes from spoken utterances in psychotherapy conversations [Website][PDF]
Karan Singla, Zhuohao Chen, David Atkins, and Shrikanth Narayanan 6:00–7:00

Spoken language understanding tasks usually rely on pipelines involving complex processing blocks such as voice activity detection, speaker diarization and Automatic speech recognition (ASR). We propose a novel framework for predicting utterance level labels directly from speech features, thus removing the dependency on first generating transcripts, and transcription free behavioral coding. Our classifier uses a pretrained Speech-2-Vector encoder as bottleneck to generate word-level representations from speech features. This pretrained encoder learns to encode speech features for a word using an objective similar to Word2Vec. Our proposed approach just uses speech features and word segmentation information for predicting spoken utterance-level target labels. We show that our model achieves competitive results to other state-of-the-art approaches which use transcribed text for the task of predicting psychotherapy-relevant behavior codes.

Session 6B: Student Research Workshop

Embeddings of Label Components for Sequence Labeling: A Case Study of Fine-grained Named Entity Recognition [Website][PDF]
Takuma Kato, Kaori Abe, Hiroki Ouchi, Shumpei Miyawaki, Jun Suzuki, and Kentaro Inui 6:00–7:00

In general, the labels used in sequence labeling consist of different types of elements. For example, IOB-format entity labels, such as B-Person and I-Person, can be decomposed into span (B and I) and type information (Person). However, while most sequence labeling models do not consider such label components, the shared components across labels, such as Person, can be beneficial for label prediction. In this work, we propose to integrate label component information as embeddings into models. Through experiments on English and Japanese fine-grained named entity recognition, we demonstrate that the proposed method improves performance, especially for instances with low-frequency labels.

Building a Japanese Typo Dataset from Wikipedia's Revision History [Website][PDF]
Yu Tanaka, Yugo Murawaki, Daisuke Kawahara, and Sadao Kurohashi 6:00–7:00

User generated texts contain many typos for which correction is necessary for NLP systems to work. Although a large number of typo—correction pairs are needed to develop a data-driven typo correction system, no such dataset is available for Japanese. In this paper, we extract over half a million Japanese typo—correction pairs from Wikipedia's revision history. Unlike other languages, Japanese poses unique challenges: (1) Japanese texts are unsegmented so that we cannot simply apply a spelling checker, and (2) the way people inputting kanji logographs results in typos with drastically different surface forms from correct ones. We address them by combining character-based extraction rules, morphological analyzers to guess readings, and various filtering methods. We evaluate the dataset using crowdsourcing and run a baseline seq2seq model for typo correction.

Preventing Critical Scoring Errors in Short Answer Scoring with Confidence Estimation [Website][PDF]
Hiroaki Funayama, Shota Sasaki, Yuichiroh Matsubayashi, Tomoya Mizumoto, Jun Suzuki, Masato Mita, and Kentaro Inui 6:00–7:00

Many recent Short Answer Scoring (SAS) systems have employed Quadratic Weighted Kappa (QWK) as the evaluation measure of their systems. However, we hypothesize that QWK is unsatisfactory for the evaluation of the SAS systems when we consider measuring their effectiveness in actual usage. We introduce a new task formulation of SAS that matches the actual usage. In our formulation, the SAS systems should extract as many scoring predictions that are not critical scoring errors (CSEs). We conduct the experiments in our new task formulation and demonstrate that a typical SAS system can predict scores with zero CSE for approximately 50% of test data at maximum by filtering out low-reliablility predictions on the basis of a certain confidence estimation. This result directly indicates the possibility of reducing half the scoring cost of human raters, which is more preferable for the evaluation of SAS systems.

Demo Session 1C

Time: *6:30–7:15*

ConvLab-2: An Open-Source Toolkit for Building, Evaluating, and Diagnosing Dialogue Systems
[Website][PDF]
Qi Zhu, Zheng Zhang, Yan Fang, Xiang Li, Ryuichi Takanobu, Jinchao Li, Baolin Peng, Jianfeng Gao, Xiaoyan Zhu, and Minlie Huang

We present ConvLab-2, an open-source toolkit that enables researchers to build task-oriented dialogue systems with state-of-the-art models, perform an end-to-end evaluation, and diagnose the weakness of systems. As the successor of ConvLab, ConvLab-2 inherits ConvLab's framework but integrates more powerful dialogue models and supports more datasets. Besides, we have developed an analysis tool and an interactive tool to assist researchers in diagnosing dialogue systems. The analysis tool presents rich statistics and summarizes common mistakes from simulated dialogues, which facilitates error analysis and system improvement. The interactive tool provides an user interface that allows developers to diagnose an assembled dialogue system by interacting with the system and modifying the output of each system component.

Demo Session 2A

Time: *8:00–8:45*

OpusFilter: A Configurable Parallel Corpus Filtering Toolbox [Website][PDF]
Mikko Aulamo, Sami Virpioja, and Jörg Tiedemann

This paper introduces OpusFilter, a flexible and modular toolbox for filtering parallel corpora. It implements a number of components based on heuristic filters, language identification libraries, character-based language models, and word alignment tools, and it can easily be extended with custom filters. Bitext segments can be ranked according to their quality or domain match using single features or a logistic regression model that can be trained without manually labeled training data. We demonstrate the effectiveness of OpusFilter on the example of a Finnish-English news translation task based on noisy web-crawled training data. Applying our tool leads to improved translation quality while significantly reducing the size of the training data, also clearly outperforming an alternative ranking given in the crawled data set. Furthermore, we show the ability of OpusFilter to perform data selection for domain adaptation.

Session 7A Overview – Tuesday, July 7, 2020 8:00–9:00

Track A *Computational Social Science and Social Media-5* Abstracts	Dynamic Online Conversation Recommendation *Zeng, Li, Wang, Mao, and Wong* [Website][PDF]	Neural Temporal Opinion Modelling for Opinion Prediction on Twitter *Zhu, He, and Zhou* [Website][PDF]	Stock Embeddings Acquired from News Articles and Price History, and an Application to Portfolio Optimization *Du and Tanaka-Ishii* [Website][PDF]	What Was Written vs. Who Read It: News Media Profiling Using Text Analysis and Social Media Context *Baly, Karadzhov, An, Kwak, Dinkov, Ali, Glass, and Nakov* [Website][PDF]	It Takes Two to Lie: One to Lie, and One to Listen *Peskov, Cheng, Elgohary, Barrow, Danescu-Niculescu-Mizil, and Boyd-Graber* [Website][PDF]
Track B *Generation-9* Abstracts	Learning Implicit Text Generation via Feature Matching *Padhi, Dognin, Bai, Nogueira dos Santos, Chenthamarakshan, Mroueh, and Das* [Website][PDF]	Two Birds, One Stone: A Simple, Unified Model for Text Generation from Structured and Unstructured Data *Shahidi, Li, and Lin* [Website][PDF]			
Track C *Machine Learning for NLP-7* Abstracts	Bayesian Hierarchical Words Representation Learning *Barkan, Rejwan, Caciularu, and Koenigstein* [Website][PDF]	How Does Selective Mechanism Improve Self-Attention Networks? *Geng, Wang, Wang, Qin, Liu, and Tu* [Website][PDF]	Pre-training Is (Almost) All You Need: An Application to Commonsense Reasoning *Tamborrino, Pellicanò, Pannier, Voitot, and Naudin* [Website][PDF]	SEEK: Segmented Embedding of Knowledge Graphs *Xu, Zheng, He, Shao, Yin, and Liu* [Website][PDF]	Zero-shot Text Classification via Reinforced Self-training *Ye, Geng, Chen, Chen, Xu, Zheng, Wang, Zhang, and Chen* [Website][PDF]
Track D *Machine Translation-9* Abstracts	A Novel Graph-based Multi-modal Fusion Encoder for Neural Machine Translation *Yin, Meng, Su, Zhou, Yang, Zhou, and Luo* [Website][PDF]	[CL] A Systematic Study of Inner-Attention-Based Sentence Representations in Multilingual Neural Machine Translation *Vázquez, Raganato, Creutz, and Tiedemann* [Website][PDF]	[TACL] Better Document-level Machine Translation with Bayes' Rule *Yu, Sartran, Stokowiec, Ling, Kong, Blunsom, and Dyer* [Website]	Dynamic Programming Encoding for Subword Segmentation in Neural Machine Translation *He, Haffari, and Norouzi* [Website][PDF]	On the Inference Calibration of Neural Machine Translation *Wang, Tu, Shi, and Liu* [Website][PDF]
	Selecting Back-translated Data from Multiple Sources for Improved Neural Machine Translation *Soto, Shterionov, Poncelas, and Way* [Website][PDF]	Successfully Applying the Stabilized Lottery Ticket Hypothesis to the Transformer Architecture *Brix, Bahar, and Ney* [Website][PDF]			
Track E *Question Answering-4* Abstracts	A Self-Training Method for Machine Reading Comprehension with Soft Evidence Extraction *Niu, Jiao, Zhou, Yao, and Huang* [Website][PDF]	Graph-to-Tree Learning for Solving Math Word Problems *Zhang, Wang, Lee, Bin, Wang, Shao, and Lim* [Website][PDF]			

Track F *Resources and Evaluation-6* Abstracts	An Effectiveness Metric for Ordinal Classification: Formal Properties and Experimental Results *Amigo, Gonzalo, Mizzaro, and Carrillo-de-Albornoz* [Website][PDF]	GLUECoS: An Evaluation Benchmark for Code-Switched NLP *Khanuja, Dandapat, Srinivasan, Sitaram, and Choudhury* [Website][PDF]			
Track G *Lexical-5* Abstracts	Adaptive Compression of Word Embeddings *Kim, Kim, and Lee* [Website][PDF]	Analysing Lexical Semantic Change with Contextualised Word Representations *Giulianelli, Del Tredici, and Fernández* [Website][PDF]	Autoencoding Keyword Correlation Graph for Document Clustering *Chiu, Sahu, Thomas, Sengupta, and Mahdy* [Website][PDF]	Autoencoding Pixies: Amortised Variational Inference with Graph Convolutions for Functional Distributional Semantics *Emerson* [Website][PDF]	BERTRAM: Improved Word Embeddings Have Big Impact on Contextualized Model Performance *Schick and Schütze* [Website][PDF]
	BiRRE: Learning Bidirectional Residual Relation Embeddings for Supervised Hypernymy Detection *Wang and HE* [Website][PDF]	Biomedical Entity Representations with Synonym Marginalization *Sung, Jeon, Lee, and Kang* [Website][PDF]	CluBERT: A Cluster-Based Approach for Learning Sense Distributions in Multiple Languages *Pasini, Scozzafava, and Scarlini* [Website][PDF]		
Track H *Sentiment Analysis, Stylistic Analysis, and Argument Mining-3* Abstracts	Adversarial and Domain-Aware BERT for Cross-Domain Sentiment Analysis *Du, Sun, Wang, Qi, and Liao* [Website][PDF]	Analyzing the Persuasive Effect of Style in News Editorial Argumentation *El Baff, Wachsmuth, Al Khatib, and Stein* [Website][PDF]	Effective Inter-Clause Modeling for End-to-End Emotion-Cause Pair Extraction *Wei, Zhao, and Mao* [Website][PDF]	Enhancing Cross-target Stance Detection with Transferable Semantic-Emotion Knowledge *Zhang, Yang, Li, Ye, Xu, and Dai* [Website][PDF]	From Arguments to Key Points: Towards Automatic Argument Summarization *Bar-Haim, Eden, Friedman, Kantor, Lahav, and Slonim* [Website][PDF]
	GoEmotions: A Dataset of Fine-Grained Emotions *Demszky, Movshovitz-Attias, Ko, Cowen, Nemade, and Ravi* [Website][PDF]	He said "who's gonna take care of your children when you are at ACL?": Reported Sexist Acts are Not Sexist *Chiril, MORICEAU, Benamara, Mari, Origgi, and Coulomb-Gully* [Website][PDF]	KinGDOM: Knowledge-Guided DOMain Adaptation for Sentiment Analysis *Ghosal, Hazarika, Roy, Majumder, Mihalcea, and Poria* [Website][PDF]	Modelling Context and Syntactical Features for Aspect-based Sentiment Analysis *Phan and Ogunbona* [Website][PDF]	Parallel Data Augmentation for Formality Style Transfer *Zhang, Ge, and SUN* [Website][PDF]
	SKEP: Sentiment Knowledge Enhanced Pre-training for Sentiment Analysis *Tian, Gao, Xiao, Liu, He, Wu, Wang, and* [Website][PDF]	SpanMlt: A Span-based Multi-Task Learning Framework for Pairwise Aspect and Opinion Terms Extraction *Zhao, Huang, Zhang, Lu, and* [Website][PDF]	[TACL] Target-Guided Structured Attention Network for Target-dependent Sentiment Analysis *Zhang, Chen, Liu, He, and Leung* [Website][PDF]	Towards Better Non-Tree Argument Mining: Proposition-Level Biaffine Parsing with Task-Specific Parameterization *Morio, Ozaki, Morishita, Koreeda, and Yanai* [Website][PDF]	

Track I *Student Research Workshop* Abstracts	Unsupervised Paraphasia Classification in Aphasic Speech *Pai, Sachdeva, Sachdeva, and Shah* [Website][PDF]	Reflection-based Word Attribute Transfer *Ishibashi, Sudoh, Yoshino, and Nakamura* [Website][PDF]	To compress or not to compress? A Finite-State approach to Nen verbal morphology *Muradoglu, Evans, and Suominen* [Website][PDF]	Embeddings of Label Components for Sequence Labeling: A Case Study of Fine-grained Named Entity Recognition *Kato, Abe, Ouchi, Miyawaki, Suzuki, and Inui* [Website][PDF]	
Track J *Tagging, Chunking and Parsing-2* Abstracts	[TACL] A Graph-based Model for Joint Chinese Word Segmentation and Dependency Parsing *Yan, Qiu, and Huang* [Website][PDF]	[CL] Abstract Syntax as Interlingua: Scaling Up the Grammatical Framework from Controlled Languages to Robust Pipelines *Ranta, Angelov, Gruzitis, and Kolachina* [Website][PDF]	An Empirical Comparison of Unsupervised Constituency Parsing Methods *Li, Cao, Cai, Jiang, and Tu* [Website][PDF]	Do Neural Language Models Show Preferences for Syntactic Formalisms? *Kulmizev, Ravishankar, Abdou, and Nivre* [Website][PDF]	Efficient Constituency Parsing by Pointing *Nguyen, Nguyen, Joty, and Li* [Website][PDF]
	Efficient Second-Order TreeCRF for Neural Dependency Parsing *Zhang, Li, and Zhang* [Website][PDF]	Enriched In-Order Linearization for Faster Sequence-to-Sequence Constituent Parsing *Fernández-González and Gómez-Rodríguez* [Website][PDF]	Exact yet Efficient Graph Parsing, Bi-directional Locality and the Constructivist Hypothesis *Ye and Sun* [Website][PDF]	Max-Margin Incremental CCG Parsing *Stanojević and Steedman* [Website][PDF]	Neural Reranking for Dependency Parsing: An Evaluation *Do and Rehbein* [Website][PDF]
	Representations of Syntax [MASK] Useful: Effects of Constituency and Dependency Structure in Recursive LSTMs *Lepori, Linzen, and McCoy* [Website][PDF]	Structure-Level Knowledge Distillation For Multilingual Sequence Labeling *Wang, Jiang, Bach, Wang, Huang, and Tu* [Website][PDF]			

Session 7A Details

Session 7A: Computational Social Science and Social Media-5

Dynamic Online Conversation Recommendation [Website][PDF]
Xingshan Zeng, Jing Li, Lu Wang, Zhiming Mao, and Kam-Fai Wong 8:00–9:00

Trending topics in social media content evolve over time, and it is therefore crucial to understand social media users and their interpersonal communications in a dynamic manner. Here we study dynamic online conversation recommendation, to help users engage in conversations that satisfy their evolving interests. While most prior work assumes static user interests, our model is able to capture the temporal aspects of user interests, and further handle future conversations that are unseen during training time. Concretely, we propose a neural architecture to exploit changes of user interactions and interests over time, to predict which discussions they are likely to enter. We conduct experiments on large-scale collections of Reddit conversations, and results on three subreddits show that our model significantly outperforms state-of-the-art models that make a static assumption of user interests. We further evaluate on handling "cold start", and observe consistently better performance by our model when considering various degrees of sparsity of user's chatting history and conversation contexts. Lastly, analyses on our model outputs indicate user interest change, explaining the advantage and efficacy of our approach.

Neural Temporal Opinion Modelling for Opinion Prediction on Twitter [Website][PDF]
Lixing Zhu, Yulan He, and Deyu Zhou 8:00–9:00

Opinion prediction on Twitter is challenging due to the transient nature of tweet content and neighbourhood context. In this paper, we model users' tweet posting behaviour as a temporal point process to jointly predict the posting time and the stance label of the next tweet given a user's historical tweet sequence and tweets posted by their neighbours. We design a topic-driven attention mechanism to capture the dynamic topic shifts in the neighbourhood context. Experimental results show that the proposed model predicts both the posting time and the stance labels of future tweets more accurately compared to a number of competitive baselines.

Stock Embeddings Acquired from News Articles and Price History, and an Application to Portfolio Optimization [Website][PDF]
Xin Du and Kumiko Tanaka-Ishii 8:00–9:00

Previous works that integrated news articles to better process stock prices used a variety of neural networks to predict price movements. The textual and price information were both encoded in the neural network, and it is therefore difficult to apply this approach in situations other than the original framework of the notoriously hard problem of price prediction. In contrast, this paper presents a method to encode the influence of news articles through a vector representation of stocks called a *stock embedding*. The stock embedding is acquired with a deep learning framework using both news articles and price history. Because the embedding takes the operational form of a vector, it is applicable to other financial problems besides price prediction. As one example application, we show the results of portfolio optimization using Reuters & Bloomberg headlines, producing a capital gain 2.8 times larger than that obtained with a baseline method using only stock price data. This suggests that the proposed stock embedding can leverage textual financial semantics to solve financial prediction problems.

What Was Written vs. Who Read It: News Media Profiling Using Text Analysis and Social Media Context [Website][PDF]
Ramy Baly, Georgi Karadzhov, Jisun An, Haewoon Kwak, Yoan Dinkov, Ahmed Ali, James Glass, and Preslav Nakov 8:00–9:00

Predicting the political bias and the factuality of reporting of entire news outlets are critical elements of media profiling, which is an understudied but an increasingly important research direction. The present level of proliferation of fake, biased, and propagandistic content online has made it impossible to fact-check every single suspicious claim, either manually or automatically. Thus, it has been proposed to profile entire news outlets and to look for those that are likely to publish fake or biased content. This makes it possible to detect likely "fake news" the moment they are published, by simply checking the reliability of their source. From a practical perspective, political bias and factuality of reporting have a linguistic aspect but also a social context. Here, we study the impact of both, namely (i) what was written (i.e., what was published by the target medium, and how it describes itself in Twitter) vs. (ii) who reads it (i.e., analyzing the target medium's audience on social media). We further study (iii) what was written about the target medium (in Wikipedia). The evaluation results show that what was written matters most, and we further show that putting all information sources together yields huge improvements over the current state-of-the-art.

It Takes Two to Lie: One to Lie, and One to Listen [Website][PDF]
Denis Peskov, Benny Cheng, Ahmed Elgohary, Joe Barrow, Cristian Danescu-Niculescu-Mizil, and Jordan Boyd-Graber 8:00–9:00

Trust is implicit in many online text conversations—striking up new friendships, or asking for tech support. But trust can be betrayed through deception. We study the language and dynamics of deception in the negotiation-based game Diplomacy, where seven players compete for world domination by forging and breaking alliances with each other. Our study with players from the Diplomacy community gathers 17,289 messages annotated by the sender for their intended truthfulness and by the receiver for their perceived truthfulness. Unlike existing datasets, this captures deception in long-lasting relationships, where the interlocutors strategically combine truth with lies to advance ob-

jectives. A model that uses power dynamics and conversational contexts can predict when a lie occurs nearly as well as human players.

Session 7A: Generation-9

Learning Implicit Text Generation via Feature Matching [Website][PDF]
Inkit Padhi, Pierre Dognin, Ke Bai, Cícero Nogueira dos Santos, Vijil Chenthamarakshan, Youssef Mroueh, and Payel Das 8:00–9:00

Generative feature matching network (GFMN) is an approach for training state-of-the-art implicit generative models for images by performing moment matching on features from pre-trained neural networks. In this paper, we present new GFMN formulations that are effective for sequential data. Our experimental results show the effectiveness of the proposed method, SeqGFMN, for three distinct generation tasks in English: unconditional text generation, class-conditional text generation, and unsupervised text style transfer. SeqGFMN is stable to train and outperforms various adversarial approaches for text generation and text style transfer.

Two Birds, One Stone: A Simple, Unified Model for Text Generation from Structured and Unstructured Data [Website][PDF]
Hamidreza Shahidi, Ming Li, and Jimmy Lin 8:00–9:00

A number of researchers have recently questioned the necessity of increasingly complex neural network (NN) architectures. In particular, several recent papers have shown that simpler, properly tuned models are at least competitive across several NLP tasks. In this work, we show that this is also the case for text generation from structured and unstructured data. We consider neural table-to-text generation and neural question generation (NQG) tasks for text generation from structured and unstructured data, respectively. Table-to-text generation aims to generate a description based on a given table, and NQG is the task of generating a question from a given passage where the generated question can be answered by a certain sub-span of the passage using NN models. Experimental results demonstrate that a basic attention-based seq2seq model trained with the exponential moving average technique achieves the state of the art in both tasks. Code is available at https://github.com/h-shahidi/2birds-gen.

Session 7A: Machine Learning for NLP-7

Bayesian Hierarchical Words Representation Learning [Website][PDF]
Oren Barkan, Idan Rejwan, Avi Caciularu, and Noam Koenigstein 8:00–9:00

This paper presents the Bayesian Hierarchical Words Representation (BHWR) learning algorithm. BHWR facilitates Variational Bayes word representation learning combined with semantic taxonomy modeling via hierarchical priors. By propagating relevant information between related words, BHWR utilizes the taxonomy to improve the quality of such representations. Evaluation of several linguistic datasets demonstrates the advantages of BHWR over suitable alternatives that facilitate Bayesian modeling with or without semantic priors. Finally, we further show that BHWR produces better representations for rare words.

How Does Selective Mechanism Improve Self-Attention Networks? [Website][PDF]
Xinwei Geng, Longyue Wang, Xing Wang, Bing Qin, Ting Liu, and Zhaopeng Tu 8:00–9:00

Self-attention networks (SANs) with selective mechanism has produced substantial improvements in various NLP tasks by concentrating on a subset of input words. However, the underlying reasons for their strong performance have not been well explained. In this paper, we bridge the gap by assessing the strengths of selective SANs (SSANs), which are implemented with a flexible and universal Gumbel-Softmax. Experimental results on several representative NLP tasks, including natural language inference, semantic role labelling, and machine translation, show that SSANs consistently outperform the standard SANs. Through well-designed probing experiments, we empirically validate that the improvement of SSANs can be attributed in part to mitigating two commonly-cited weaknesses of SANs: word order encoding and structure modeling. Specifically, the selective mechanism improves SANs by paying more attention to content words that contribute to the meaning of the sentence.

Pre-training Is (Almost) All You Need: An Application to Commonsense Reasoning [Website][PDF]
Alexandre Tamborrino, Nicola Pellicanò, Baptiste Pannier, Pascal Voitot, and Louise Naudin 8:00–9:00

Fine-tuning of pre-trained transformer models has become the standard approach for solving common NLP tasks. Most of the existing approaches rely on a randomly initialized classifier on top of such networks. We argue that this fine-tuning procedure is sub-optimal as the pre-trained model has no prior on the specific classifier labels, while it might have already learned an intrinsic textual representation of the task. In this paper, we introduce a new scoring method that casts a plausibility ranking task in a full-text format and leverages the masked language modeling head tuned during the pre-training phase. We study commonsense reasoning tasks where the model must rank a set of hypotheses given a premise, focusing on the COPA, Swag, HellaSwag and CommonsenseQA datasets. By exploiting our scoring method without fine-tuning, we are able to produce strong baselines (e.g. 80% test accuracy on COPA) that are comparable to supervised approaches. Moreover, when fine-tuning directly on the proposed scoring function, we show that our method provides a much more stable training phase across random restarts (e.g x10 standard deviation reduction on COPA test accuracy) and requires less annotated data than the standard classifier approach to reach equivalent performances.

SEEK: Segmented Embedding of Knowledge Graphs [Website][PDF]
Wentao Xu, Shun Zheng, Liang He, Bin Shao, Jian Yin, and Tie-Yan Liu 8:00–9:00

In recent years, knowledge graph embedding becomes a pretty hot research topic of artificial intelligence and plays increasingly vital roles in various downstream applications, such as recommendation and question answering. However, existing methods for knowledge graph embedding can not make a proper trade-off between the model complexity and the model expressiveness, which makes them still far from satisfactory. To mitigate this problem, we propose a lightweight modeling framework that can achieve highly competitive relational expressiveness without increasing the model complexity. Our framework focuses on the design of scoring functions and highlights two critical characteristics: 1) facilitating sufficient feature interactions; 2) preserving both symmetry and antisymmetry properties of relations. It is noteworthy that owing to the general and elegant design of scoring functions, our framework can incorporate many famous existing methods as special cases. Moreover, extensive experiments on public benchmarks demonstrate the efficiency and effectiveness of our framework. Source codes and data can be found at `https://github.com/Wentao-Xu/SEEK`.

Zero-shot Text Classification via Reinforced Self-training [Website][PDF]
Zhiquan Ye, Yuxia Geng, Jiaoyan Chen, Jingmin Chen, Xiaoxiao Xu, SuHang Zheng, Feng Wang, Jun Zhang, and Huajun Chen 8:00–9:00

Zero-shot learning has been a tough problem since no labeled data is available for unseen classes during training, especially for classes with low similarity. In this situation, transferring from seen classes to unseen classes is extremely hard. To tackle this problem, in this paper we propose a self-training based method to efficiently leverage unlabeled data. Traditional self-training methods use fixed heuristics to select instances from unlabeled data, whose performance varies among different datasets. We propose a reinforcement learning framework to learn data selection strategy automatically and provide more reliable selection. Experimental results on both benchmarks and a real-world e-commerce dataset show that our approach significantly outperforms previous methods in zero-shot text classification

Session 7A: Machine Translation-9

A Novel Graph-based Multi-modal Fusion Encoder for Neural Machine Translation [Website][PDF]
Yongjing Yin, Fandong Meng, Jinsong Su, Chulun Zhou, Zhengyuan Yang, Jie Zhou, and Jiebo Luo 8:00–9:00

Multi-modal neural machine translation (NMT) aims to translate source sentences into a target language paired with images. However, dominant multi-modal NMT models do not fully exploit fine-grained semantic correspondences between semantic units of different modalities, which have potential to refine multi-modal representation learning. To deal with this issue, in this paper, we propose a novel graph-based multi-modal fusion encoder for NMT. Specifically, we first represent the input sentence and image using a unified multi-modal graph, which captures various semantic relationships between multi-modal semantic units (words and visual objects). We then stack multiple graph-based multi-modal fusion layers that iteratively perform semantic interactions to learn node representations. Finally, these representations provide an attention-based context vector for the decoder. We evaluate our proposed encoder on the Multi30K datasets. Experimental results and in-depth analysis show the superiority of our multi-modal NMT model.

[CL] A Systematic Study of Inner-Attention-Based Sentence Representations in Multilingual Neural Machine Translation [Website][PDF]
Raúl Vázquez, Alessandro Raganato, Mathias Creutz, and Jörg Tiedemann 8:00–9:00

Neural machine translation has considerably improved the quality of automatic translations by learning good representations of input sentences. In this article, we explore a multilingual translation model capable of producing fixed-size sentence representations by incorporating an intermediate crosslingual shared layer, which we refer to as attention bridge. This layer exploits the semantics from each language and develops into a language-agnostic meaning representation that can be efficiently used for transfer learning.We systematically study the impact of the size of the attention bridge and the effect of including additional languages in the model. In contrast to related previous work, we demonstrate that there is no conflict between translation performance and the use of sentence representations in downstream tasks. In particular, we show that larger intermediate layers not only improve translation quality, especially for long sentences, but also push the accuracy of trainable classification tasks. Nevertheless, shorter representations lead to increased compression that is beneficial in non-trainable similarity tasks. Similarly, we show that trainable downstream tasks benefit from multilingual models, whereas additional language signals do not improve performance in non-trainable benchmarks. This is an important insight that helps to properly design models for specific applications. Finally, we also include an in-depth analysis of the proposed attention bridge and its ability of encoding linguistic properties. We carefully analyze the information that is captured by individual attention heads and identify interesting patterns that explain the performance of specific settings in linguistic probing tasks.

[TACL] Better Document-level Machine Translation with Bayes' Rule [Website]
Lei Yu, Laurent Sartran, Wojciech Stokowiec, Wang Ling, Lingpeng Kong, Phil Blunsom, and Chris Dyer 8:00–9:00

We show that Bayes' rule provides an effective mechanism for creating document translation models that can be learned from only parallel sentences and monolingual documents—a compelling benefit as parallel documents are not always available. In our formulation, the posterior probability of a candidate translation is the product of the unconditional (prior) probability of the candidate output document and the "reverse translation probability" of translating the candidate output back into the source language. Our proposed model uses a powerful autoregressive language model as the prior on target language documents, but it assumes that each sentence is translated independently from the target to the source language. Crucially, at test time, when a source document is observed, the document language model prior induces dependencies between the translations of the source sentences in the posterior. The model's independence assumption not only enables efficient use of available data, but it additionally admits a practical left-to-right beam-search algorithm for carrying out inference. Experiments show that our model benefits from using cross-sentence context in the language model, and it outperforms existing document translation approaches.

Dynamic Programming Encoding for Subword Segmentation in Neural Machine Translation [Website][PDF]
Xuanli He, Gholamreza Haffari, and Mohammad Norouzi 8:00–9:00

This paper introduces Dynamic Programming Encoding (DPE), a new segmentation algorithm for tokenizing sentences into subword units. We view the subword segmentation of output sentences as a latent variable that should be marginalized out for learning and inference. A mixed character-subword transformer is proposed, which enables exact log marginal likelihood estimation and exact MAP inference to find target segmentations with maximum posterior probability. DPE uses a lightweight mixed character-subword transformer as a means of pre-processing parallel data to segment output sentences using dynamic programming. Empirical results on machine translation suggest that DPE is effective for segmenting output sentences and can be combined with BPE dropout for stochastic segmentation of source sentences. DPE achieves an average improvement of 0.9 BLEU over BPE (Sennrich et al., 2016) and an average improvement of 0.55 BLEU over BPE dropout (Provilkov et al., 2019) on several WMT datasets including English <=> (German, Romanian, Estonian, Finnish, Hungarian).

On the Inference Calibration of Neural Machine Translation [Website][PDF]
Shuo Wang, Zhaopeng Tu, Shuming Shi, and Yang Liu 8:00–9:00

Confidence calibration, which aims to make model predictions equal to the true correctness measures, is important for neural machine translation (NMT) because it is able to offer useful indicators of translation errors in the generated output. While prior studies have shown that NMT models trained with label smoothing are well-calibrated on the

ground-truth training data, we find that miscalibration still remains a severe challenge for NMT during inference due to the discrepancy between training and inference. By carefully designing experiments on three language pairs, our work provides in-depth analyses of the correlation between calibration and translation performance as well as linguistic properties of miscalibration and reports a number of interesting findings that might help humans better analyze, understand and improve NMT models. Based on these observations, we further propose a new graduated label smoothing method that can improve both inference calibration and translation performance.

Selecting Backtranslated Data from Multiple Sources for Improved Neural Machine Translation [Website][PDF]

Xabier Soto, Dimitar Shterionov, Alberto Poncelas, and Andy Way 8:00–9:00

Machine translation (MT) has benefited from using synthetic training data originating from translating monolingual corpora, a technique known as backtranslation. Combining backtranslated data from different sources has led to better results than when using such data in isolation. In this work we analyse the impact that data translated with rule-based, phrase-based statistical and neural MT systems has on new MT systems. We use a real-world low-resource use-case (Basque-to-Spanish in the clinical domain) as well as a high-resource language pair (German-to-English) to test different scenarios with backtranslation and employ data selection to optimise the synthetic corpora. We exploit different data selection strategies in order to reduce the amount of data used, while at the same time maintaining high-quality MT systems. We further tune the data selection method by taking into account the quality of the MT systems used for backtranslation and lexical diversity of the resulting corpora. Our experiments show that incorporating backtranslated data from different sources can be beneficial, and that availing of data selection can yield improved performance.

Successfully Applying the Stabilized Lottery Ticket Hypothesis to the Transformer Architecture [Website][PDF]

Christopher Brix, Parnia Bahar, and Hermann Ney 8:00–9:00

Sparse models require less memory for storage and enable a faster inference by reducing the necessary number of FLOPs. This is relevant both for time-critical and on-device computations using neural networks. The stabilized lottery ticket hypothesis states that networks can be pruned after none or few training iterations, using a mask computed based on the unpruned converged model. On the transformer architecture and the WMT 2014 English-to-German and English-to-French tasks, we show that stabilized lottery ticket pruning performs similar to magnitude pruning for sparsity levels of up to 85%, and propose a new combination of pruning techniques that outperforms all other techniques for even higher levels of sparsity. Furthermore, we confirm that the parameter's initial sign and not its specific value is the primary factor for successful training, and show that magnitude pruning cannot be used to find winning lottery tickets.

Session 7A: Question Answering-4

A Self-Training Method for Machine Reading Comprehension with Soft Evidence Extraction [Website][PDF]
Yilin Niu, Fangkai Jiao, Mantong Zhou, Ting Yao, jingfang xu jingfang, and Minlie Huang 8:00–9:00

Neural models have achieved great success on machine reading comprehension (MRC), many of which typically consist of two components: an evidence extractor and an answer predictor. The former seeks the most relevant information from a reference text, while the latter is to locate or generate answers from the extracted evidence. Despite the importance of evidence labels for training the evidence extractor, they are not cheaply accessible, particularly in many non-extractive MRC tasks such as YES/NO question answering and multi-choice MRC. To address this problem, we present a Self-Training method (STM), which supervises the evidence extractor with auto-generated evidence labels in an iterative process. At each iteration, a base MRC model is trained with golden answers and noisy evidence labels. The trained model will predict pseudo evidence labels as extra supervision in the next iteration. We evaluate STM on seven datasets over three MRC tasks. Experimental results demonstrate the improvement on existing MRC models, and we also analyze how and why such a self-training method works in MRC.

Graph-to-Tree Learning for Solving Math Word Problems [Website][PDF]
Jipeng Zhang, Lei Wang, Roy Ka-Wei Lee, Yi Bin, Yan Wang, Jie Shao, and Ee-Peng Lim 8:00–9:00

While the recent tree-based neural models have demonstrated promising results in generating solution expression for the math word problem (MWP), most of these models do not capture the relationships and order information among the quantities well. This results in poor quantity representations and incorrect solution expressions. In this paper, we propose Graph2Tree, a novel deep learning architecture that combines the merits of the graph-based encoder and tree-based decoder to generate better solution expressions. Included in our Graph2Tree framework are two graphs, namely the Quantity Cell Graph and Quantity Comparison Graph, which are designed to address limitations of existing methods by effectively representing the relationships and order information among the quantities in MWPs. We conduct extensive experiments on two available datasets. Our experiment results show that Graph2Tree outperforms the state-of-the-art baselines on two benchmark datasets significantly. We also discuss case studies and empirically examine Graph2Tree's effectiveness in translating the MWP text into solution expressions.

Session 7A: Resources and Evaluation-6

An Effectiveness Metric for Ordinal Classification: Formal Properties and Experimental Results [Website][PDF]
Enrique Amigo, Julio Gonzalo, Stefano Mizzaro, and Jorge Carrillo-de-Albornoz 8:00–9:00

In Ordinal Classification tasks, items have to be assigned to classes that have a relative ordering, such as "positive", "neutral", "negative" in sentiment analysis. Remarkably, the most popular evaluation metrics for ordinal classification tasks either ignore relevant information (for instance, precision/recall on each of the classes ignores their relative ordering) or assume additional information (for instance, Mean Average Error assumes absolute distances between classes). In this paper we propose a new metric for Ordinal Classification, Closeness Evaluation Measure, that is rooted on Measurement Theory and Information Theory. Our theoretical analysis and experimental results over both synthetic data and data from NLP shared tasks indicate that the proposed metric captures quality aspects from different traditional tasks simultaneously. In addition, it generalizes some popular classification (nominal scale) and error minimization (interval scale) metrics, depending on the measurement scale in which it is instantiated.

GLUECoS: An Evaluation Benchmark for Code-Switched NLP [Website][PDF]
Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury 8:00–9:00

Code-switching is the use of more than one language in the same conversation or utterance. Recently, multilingual contextual embedding models, trained on multiple monolingual corpora, have shown promising results on cross-lingual and multilingual tasks. We present an evaluation benchmark, GLUECoS, for code-switched languages, that spans several NLP tasks in English-Hindi and English-Spanish. Specifically, our evaluation benchmark includes Language Identification from text, POS tagging, Named Entity Recognition, Sentiment Analysis, Question Answering and a new task for code-switching, Natural Language Inference. We present results on all these tasks using cross-lingual word embedding models and multilingual models. In addition, we fine-tune multilingual models on artificially generated code-switched data. Although multilingual models perform significantly better than cross-lingual models, our results show that in most tasks, across both language pairs, multilingual models fine-tuned on code-switched data perform best, showing that multilingual models can be further optimized for code-switching tasks.

Session 7A Semantics: Lexical-5

Adaptive Compression of Word Embeddings [Website][PDF]
Yeachan Kim, Kang-Min Kim, and SangKeun Lee 8:00–9:00

Distributed representations of words have been an indispensable component for natural language processing (NLP) tasks. However, the large memory footprint of word embeddings makes it challenging to deploy NLP models to memory-constrained devices (e.g., self-driving cars, mobile devices). In this paper, we propose a novel method to adaptively compress word embeddings. We fundamentally follow a code-book approach that represents words as discrete codes such as (8, 5, 2, 4). However, unlike prior works that assign the same length of codes to all words, we adaptively assign different lengths of codes to each word by learning downstream tasks. The proposed method works in two steps. First, each word directly learns to select its code length in an end-to-end manner by applying the Gumbel-softmax tricks. After selecting the code length, each word learns discrete codes through a neural network with a binary constraint. To showcase the general applicability of the proposed method, we evaluate the performance on four different downstream tasks. Comprehensive evaluation results clearly show that our method is effective and makes the highly compressed word embeddings without hurting the task accuracy. Moreover, we show that our model assigns word to each code-book by considering the significance of tasks.

Analysing Lexical Semantic Change with Contextualised Word Representations [Website][PDF]
Mario Giulianelli, Marco Del Tredici, and Raquel Fernández 8:00–9:00

This paper presents the first unsupervised approach to lexical semantic change that makes use of contextualised word representations. We propose a novel method that exploits the BERT neural language model to obtain representations of word usages, clusters these representations into usage types, and measures change along time with three proposed metrics. We create a new evaluation dataset and show that the model representations and the detected semantic shifts are positively correlated with human judgements. Our extensive qualitative analysis demonstrates that our method captures a variety of synchronic and diachronic linguistic phenomena. We expect our work to inspire further research in this direction.

Autoencoding Keyword Correlation Graph for Document Clustering [Website][PDF]
Billy Chiu, Sunil Kumar Sahu, Derek Thomas, Neha Sengupta, and Mohammady Mahdy 8:00–9:00

Document clustering requires a deep understanding of the complex structure of long-text; in particular, the intra-sentential (local) and inter-sentential features (global). Existing representation learning models do not fully capture these features. To address this, we present a novel graph-based representation for document clustering that builds a *graph autoencoder* (GAE) on a Keyword Correlation Graph. The graph is constructed with topical keywords as nodes and multiple local and global features as edges. A GAE is employed to aggregate the two sets of features by learning a latent representation which can jointly reconstruct them. Clustering is then performed on the learned representations, using vector dimensions as features for inducing document classes. Extensive experiments on two datasets show that the features learned by our approach can achieve better clustering performance than other existing features, including term frequency-inverse document frequency and average embedding.

Autoencoding Pixies: Amortised Variational Inference with Graph Convolutions for Functional Distributional Semantics [Website][PDF]
Guy Emerson 8:00–9:00

Functional Distributional Semantics provides a linguistically interpretable framework for distributional semantics, by representing the meaning of a word as a function (a binary classifier), instead of a vector. However, the large number of latent variables means that inference is computationally expensive, and training a model is therefore slow to converge. In this paper, I introduce the Pixie Autoencoder, which augments the generative model of Functional Distributional Semantics with a graph-convolutional neural network to perform amortised variational inference. This allows the model to be trained more effectively, achieving better results on two tasks (semantic similarity in context and semantic composition), and outperforming BERT, a large pre-trained language model.

BERTRAM: Improved Word Embeddings Have Big Impact on Contextualized Model Performance [Website][PDF]
Timo Schick and Hinrich Schütze 8:00–9:00

Pretraining deep language models has led to large performance gains in NLP. Despite this success, Schick and Schütze (2020) recently showed that these models struggle to understand rare words. For static word embeddings, this problem has been addressed by separately learning representations for rare words. In this work, we transfer this idea to pretrained language models: We introduce BERTRAM, a powerful architecture based on BERT that is capable of inferring high-quality embeddings for rare words that are suitable as input representations for deep language models. This is achieved by enabling the surface form and contexts of a word to interact with each other in a deep architecture. Integrating BERTRAM into BERT leads to large performance increases due to improved representations of rare and medium frequency words on both a rare word probing task and three downstream tasks.

BiRRE: Learning Bidirectional Residual Relation Embeddings for Supervised Hypernymy Detection [Website][PDF]
Chengyu Wang and XIAOFENG HE 8:00–9:00

The hypernymy detection task has been addressed under various frameworks. Previously, the design of unsupervised hypernymy scores has been extensively studied. In contrast, supervised classifiers, especially distributional models, leverage the global contexts of terms to make predictions, but are more likely to suffer from "lexical memorization".

In this work, we revisit supervised distributional models for hypernymy detection. Rather than taking embeddings of two terms as classification inputs, we introduce a representation learning framework named Bidirectional Residual Relation Embeddings (BiRRE). In this model, a term pair is represented by a BiRRE vector as features for hypernymy classification, which models the possibility of a term being mapped to another in the embedding space by hypernymy relations. A Latent Projection Model with Negative Regularization (LPMNR) is proposed to simulate how hypernyms and hyponyms are generated by neural language models, and to generate BiRRE vectors based on bidirectional residuals of projections. Experiments verify BiRRE outperforms strong baselines over various evaluation frameworks.

Biomedical Entity Representations with Synonym Marginalization [Website][PDF]
Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang 8:00–9:00

Biomedical named entities often play important roles in many biomedical text mining tools. However, due to the incompleteness of provided synonyms and numerous variations in their surface forms, normalization of biomedical entities is very challenging. In this paper, we focus on learning representations of biomedical entities solely based on the synonyms of entities. To learn from the incomplete synonyms, we use a model-based candidate selection and maximize the marginal likelihood of the synonyms present in top candidates. Our model-based candidates are iteratively updated to contain more difficult negative samples as our model evolves. In this way, we avoid the explicit pre-selection of negative samples from more than 400K candidates. On four biomedical entity normalization datasets having three different entity types (disease, chemical, adverse reaction), our model BioSyn consistently outperforms previous state-of-the-art models almost reaching the upper bound on each dataset.

CluBERT: A Cluster-Based Approach for Learning Sense Distributions in Multiple Languages [Website][PDF]
Tommaso Pasini, Federico Scozzafava, and Bianca Scarlini 8:00–9:00

Knowing the Most Frequent Sense (MFS) of a word has been proved to help Word Sense Disambiguation (WSD) models significantly. However, the scarcity of sense-annotated data makes it difficult to induce a reliable and high-coverage distribution of the meanings in a language vocabulary. To address this issue, in this paper we present CluBERT, an automatic and multilingual approach for inducing the distributions of word senses from a corpus of raw sentences. Our experiments show that CluBERT learns distributions over English senses that are of higher quality than those extracted by alternative approaches. When used to induce the MFS of a lemma, CluBERT attains state-of-the-art results on the English Word Sense Disambiguation tasks and helps to improve the disambiguation performance of two off-the-shelf WSD models. Moreover, our distributions also prove to be effective in other languages, beating all their alternatives for computing the MFS on the multilingual WSD tasks. We release our sense distributions in five different languages at https://github.com/SapienzaNLP/clubert.

Session 7A: Sentiment Analysis, Stylistic Analysis, and Argument Mining-3

Adversarial and Domain-Aware BERT for Cross-Domain Sentiment Analysis [Website][PDF]
Chunning Du, Haifeng Sun, Jingyu Wang, Qi Qi, and Jianxin Liao 8:00–9:00

Cross-domain sentiment classification aims to address the lack of massive amounts of labeled data. It demands to predict sentiment polarity on a target domain utilizing a classifier learned from a source domain. In this paper, we investigate how to efficiently apply the pre-training language model BERT on the unsupervised domain adaptation. Due to the pre-training task and corpus, BERT is task-agnostic, which lacks domain awareness and can not distinguish the characteristic of source and target domain when transferring knowledge. To tackle these problems, we design a post-training procedure, which contains the target domain masked language model task and a novel domain-distinguish pre-training task. The post-training procedure will encourage BERT to be domain-aware and distill the domain-specific features in a self-supervised way. Based on this, we could then conduct the adversarial training to derive the enhanced domain-invariant features. Extensive experiments on Amazon dataset show that our model outperforms state-of-the-art methods by a large margin. The ablation study demonstrates that the remarkable improvement is not only from BERT but also from our method.

Analyzing the Persuasive Effect of Style in News Editorial Argumentation [Website][PDF]
Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein 8:00–9:00

News editorials argue about political issues in order to challenge or reinforce the stance of readers with different ideologies. Previous research has investigated such persuasive effects for argumentative content. In contrast, this paper studies how important the style of news editorials is to achieve persuasion. To this end, we first compare content- and style-oriented classifiers on editorials from the liberal NYTimes with ideology-specific effect annotations. We find that conservative readers are resistant to NYTimes style, but on liberals, style even has more impact than content. Focusing on liberals, we then cluster the leads, bodies, and endings of editorials, in order to learn about writing style patterns of effective argumentation.

Effective Inter-Clause Modeling for End-to-End Emotion-Cause Pair Extraction [Website][PDF]
Penghui Wei, Jiahao Zhao, and Wenji Mao 8:00–9:00

Emotion-cause pair extraction aims to extract all emotion clauses coupled with their cause clauses from a given document. Previous work employs two-step approaches, in which the first step extracts emotion clauses and cause clauses separately, and the second step trains a classifier to filter out negative pairs. However, such pipeline-style system for emotion-cause pair extraction is suboptimal because it suffers from error propagation and the two steps may not adapt to each other well. In this paper, we tackle emotion-cause pair extraction from a ranking perspective, i.e., ranking clause pair candidates in a document, and propose a one-step neural approach which emphasizes inter-clause modeling to perform end-to-end extraction. It models the interrelations between the clauses in a document to learn clause representations with graph attention, and enhances clause pair representations with kernel-based relative position embedding for effective ranking. Experimental results show that our approach significantly outperforms the current two-step systems, especially in the condition of extracting multiple pairs in one document.

Enhancing Cross-target Stance Detection with Transferable Semantic-Emotion Knowledge [Website][PDF]
Bowen Zhang, Min Yang, Xutao Li, Yunming Ye, Xiaofei Xu, and Kuai Dai 8:00–9:00

Stance detection is an important task, which aims to classify the attitude of an opinionated text towards a given target. Remarkable success has been achieved when sufficient labeled training data is available. However, annotating sufficient data is labor-intensive, which establishes significant barriers for generalizing the stance classifier to the data with new targets. In this paper, we proposed a Semantic-Emotion Knowledge Transferring (SEKT) model for cross-target stance detection, which uses the external knowledge (semantic and emotion lexicons) as a bridge to enable knowledge transfer across different targets. Specifically, a semantic-emotion heterogeneous graph is constructed from external semantic and emotion lexicons, which is then fed into a graph convolutional network to learn multi-hop semantic connections between words and emotion tags. Then, the learned semantic-emotion graph representation, which serves as prior knowledge bridging the gap between the source and target domains, is fully integrated into the bidirectional long short-term memory (BiLSTM) stance classifier by adding a novel knowledge-aware memory unit to the BiLSTM cell. Extensive experiments on a large real-world dataset demonstrate the superiority of SEKT against the state-of-the-art baseline methods.

From Arguments to Key Points: Towards Automatic Argument Summarization [Website][PDF]
Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim 8:00–9:00

Generating a concise summary from a large collection of arguments on a given topic is an intriguing yet understudied problem. We propose to represent such summaries as a small set of talking points, termed *key points*, each scored according to its salience. We show, by analyzing a large dataset of crowd-contributed arguments, that a small number of key points per topic is typically sufficient for covering the vast majority of the arguments. Furthermore, we found that a domain expert can often predict these key points in advance. We study the task of argument-to-key point mapping, and introduce a novel large-scale dataset for this task. We report empirical results for an extensive set of experiments with this dataset, showing promising performance.

GoEmotions: A Dataset of Fine-Grained Emotions [Website][PDF]
Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi 8:00–9:00

Understanding emotion expressed in language has a wide range of applications, from building empathetic chatbots to detecting harmful online behavior. Advancement in this area can be improved using large-scale datasets with a fine-grained typology, adaptable to multiple downstream tasks. We introduce GoEmotions, the largest manually annotated dataset of 58k English Reddit comments, labeled for 27 emotion categories or Neutral. We demonstrate the high quality of the annotations via Principal Preserved Component Analysis. We conduct transfer learning experiments with existing emotion benchmarks to show that our dataset generalizes well to other domains and different emotion taxonomies. Our BERT-based model achieves an average F1-score of .46 across our proposed taxonomy, leaving much room for improvement.

He said "who's gonna take care of your children when you are at ACL?": Reported Sexist Acts are Not Sexist [Website][PDF]
Patricia Chiril, Véronique MORICEAU, Farah Benamara, Alda Mari, Gloria Origgi, and Marlène Coulomb-Gully 8:00–9:00

In a context of offensive content mediation on social media now regulated by European laws, it is important not only to be able to automatically detect sexist content but also to identify if a message with a sexist content is really sexist or is a story of sexism experienced by a woman. We propose: (1) a new characterization of sexist content inspired by speech acts theory and discourse analysis studies, (2) the first French dataset annotated for sexism detection, and (3) a set of deep learning experiments trained on top of a combination of several tweet's vectorial representations (word embeddings, linguistic features, and various generalization strategies). Our results are encouraging and constitute a first step towards offensive content moderation.

KinGDOM: Knowledge-Guided DOMain Adaptation for Sentiment Analysis [Website][PDF]
Deepanway Ghosal, Devamanyu Hazarika, Abhinaba Roy, Navonil Majumder, Rada Mihalcea, and Soujanya Poria 8:00–9:00

Cross-domain sentiment analysis has received significant attention in recent years, prompted by the need to combat the domain gap between different applications that make use of sentiment analysis. In this paper, we take a novel perspective on this task by exploring the role of external commonsense knowledge. We introduce a new framework, KinGDOM, which utilizes the ConceptNet knowledge graph to enrich the semantics of a document by providing both domain-specific and domain-general background concepts. These concepts are learned by training a graph convolutional autoencoder that leverages inter-domain concepts in a domain-invariant manner. Conditioning a popular domain-adversarial baseline method with these learned concepts helps improve its performance over state-of-the-art approaches, demonstrating the efficacy of our proposed framework.

Modelling Context and Syntactical Features for Aspect-based Sentiment Analysis [Website][PDF]
Minh Hieu Phan and Philip O. Ogunbona 8:00–9:00

The aspect-based sentiment analysis (ABSA) consists of two conceptual tasks, namely an aspect extraction and an aspect sentiment classification. Rather than considering the tasks separately, we build an end-to-end ABSA solution. Previous works in ABSA tasks did not fully leverage the importance of syntactical information. Hence, the aspect extraction model often failed to detect the boundaries of multi-word aspect terms. On the other hand, the aspect sentiment classifier was unable to account for the syntactical correlation between aspect terms and the context words. This paper explores the grammatical aspect of the sentence and employs the self-attention mechanism for syntactical learning. We combine part-of-speech embeddings, dependency-based embeddings and contextualized embeddings (e.g. BERT, RoBERTa) to enhance the performance of the aspect extractor. We also propose the syntactic relative distance to de-emphasize the adverse effects of unrelated words, having weak syntactic connection with the aspect terms. This increases the accuracy of the aspect sentiment classifier. Our solutions outperform the state-of-the-art models on SemEval-2014 dataset in both two subtasks.

Parallel Data Augmentation for Formality Style Transfer [Website][PDF]
Yi Zhang, Tao Ge, and Xu SUN 8:00–9:00

The main barrier to progress in the task of Formality Style Transfer is the inadequacy of training data. In this paper, we study how to augment parallel data and propose novel and simple data augmentation methods for this task to obtain useful sentence pairs with easily accessible models and systems. Experiments demonstrate that our augmented parallel data largely helps improve formality style transfer when it is used to pre-train the model, leading to the state-of-the-art results in the GYAFC benchmark dataset.

SKEP: Sentiment Knowledge Enhanced Pre-training for Sentiment Analysis [Website][PDF]
Hao Tian, Can Gao, Xinyan Xiao, Hao Liu, Bolei He, Hua Wu, Haifeng Wang, and feng wu feng 8:00–9:00

Recently, sentiment analysis has seen remarkable advance with the help of pre-training approaches. However, sentiment knowledge, such as sentiment words and aspect-sentiment pairs, is ignored in the process of pre-training, despite the fact that they are widely used in traditional sentiment analysis approaches. In this paper, we introduce Sentiment Knowledge Enhanced Pre-training (SKEP) in order to learn a unified sentiment representation for multiple sentiment analysis tasks. With the help of automatically-mined knowledge, SKEP conducts sentiment masking and constructs three sentiment knowledge prediction objectives, so as to embed sentiment information at the word, polarity and aspect level into pre-trained sentiment representation. In particular, the prediction of aspect-sentiment pairs is converted into multi-label classification, aiming to capture the dependency between words in a pair. Experiments on three kinds of sentiment tasks show that SKEP significantly outperforms strong pre-training baseline, and achieves new state-of-the-art results on most of the test datasets. We release our code at https://github.com/baidu/Senta.

SpanMlt: A Span-based Multi-Task Learning Framework for Pair-wise Aspect and Opinion Terms Extraction [Website][PDF]
He Zhao, Longtao Huang, Rong Zhang, Quan Lu, and hui xue hui 8:00–9:00

Aspect terms extraction and opinion terms extraction are two key problems of fine-grained Aspect Based Sentiment Analysis (ABSA). The aspect-opinion pairs can provide a global profile about a product or service for consumers and opinion mining systems. However, traditional methods can not directly output aspect-opinion pairs without given aspect terms or opinion terms. Although some recent co-extraction methods have been proposed to extract both terms jointly, they fail to extract them as pairs. To this end, this paper proposes an end-to-end method to solve the task of Pair-wise Aspect and Opinion Terms Extraction (PAOTE). Furthermore, this paper treats the problem from a perspective of joint term and relation extraction rather than under the sequence tagging formulation performed in most prior works. We propose a multi-task learning framework based on shared spans, where the terms are extracted under the supervision of span boundaries. Meanwhile, the pair-wise relations are jointly identified using the span representations. Extensive experiments show that our model consistently outperforms state-of-the-art methods.

[TACL] Target-Guided Structured Attention Network for Target-dependent Sentiment Analysis [Website][PDF]
Ji Zhang, Chengyao Chen, Pengfei Liu, Chao He, and Cane Wing-Ki Leung 8:00–9:00

Target-dependent sentiment analysis (TDSA) aims to classify the sentiment of a text towards a given target. The major challenge of this task lies in modeling the semantic relatedness between a target and its context sentence. This paper proposes a novel Target-Guided Structured Attention Network (TG-SAN), which captures target-related contexts for TDSA in a fine-to-coarse manner. Given a target and its context sentence, the proposed TG-SAN first identifies multiple semantic segments from the sentence using a target-guided structured attention mechanism. It then fuses the extracted segments based on their relatedness with the target for sentiment classification. We present comprehensive comparative experiments on three benchmarks with three major findings. Firstly, TG-SAN outperforms the state-of-the-art by up to 1.61% and 3.58% in terms of accuracy and Marco-F1 respectively. Secondly, it shows a strong advantage in determining the sentiment of a target when the context sentence contains multiple semantic segments. Lastly, the attention results produced by TG-SAN are highly interpretable as visualization results shown.

Towards Better Non-Tree Argument Mining: Proposition-Level Biaffine Parsing with Task-Specific Parameterization [Website][PDF]
Gaku Morio, Hiroaki Ozaki, Terufumi Morishita, Yuta Koreeda, and Kohsuke Yanai 8:00–9:00

State-of-the-art argument mining studies have advanced the techniques for predicting argument structures. However, the technology for capturing non-tree-structured arguments is still in its infancy. In this paper, we focus on non-tree argument mining with a neural model. We jointly predict proposition types and edges between propositions. Our proposed model incorporates (i) task-specific parameterization (TSP) that effectively encodes a sequence of propositions and (ii) a proposition-level biaffine attention (PLBA) that can predict a non-tree argument consisting of edges. Experimental results show that both TSP and PLBA boost edge prediction performance compared to baselines.

Session 7A: Student Research Workshop

Unsupervised Paraphasia Classification in Aphasic Speech [Website][PDF]
Sharan Pai, Nikhil Sachdeva, Prince Sachdeva, and Rajiv Ratn Shah 8:00–9:00

Aphasia is a speech and language disorder which results from brain damage, often characterized by word retrieval deficit (anomia) resulting in naming errors (paraphasia). Automatic paraphasia detection has many benefits for both treatment and diagnosis of Aphasia and its type. But supervised learning methods cant be properly utilized as there is a lack of aphasic speech data. In this paper, we describe our novel unsupervised method which can be implemented without the need for labeled paraphasia data. Our evaluations show that our method outperforms previous work based on supervised learning and transfer learning approaches for English. We demonstrate the utility of our method as an essential first step in developing augmentative and alternative communication (AAC) devices for patients suffering from aphasia in any language.

Reflection-based Word Attribute Transfer [Website][PDF]
Yoichi Ishibashi, Katsuhito Sudoh, Koichiro Yoshino, and Satoshi Nakamura 8:00–9:00

Word embeddings, which often represent such analogic relations as king - man + woman ~ queen, can be used to change a word's attribute, including its gender. For transferring king into queen in this analogy-based manner, we subtract a difference vector man - woman based on the knowledge that king is male. However, developing such knowledge is very costly for words and attributes. In this work, we propose a novel method for word attribute transfer based on reflection mappings without such an analogy operation. Experimental results show that our proposed method can transfer the word attributes of the given words without changing the words that do not have the target attributes.

To compress or not to compress? A Finite-State approach to Nen verbal morphology [Website][PDF]
Saliha Muradoglu, Nicholas Evans, and Hanna Suominen 8:00–9:00

This paper describes the development of a verbal morphological parser for an under-resourced Papuan language, Nen. Nen verbal morphology is particularly complex, with a transitive verb taking up to 1,740 unique features. The structural properties exhibited by Nen verbs raises interesting choices for analysis. Here we compare two possible methods of analysis: 'Chunking' and decomposition. 'Chunking' refers to the concept of collating morphological segments into one, whereas the decomposition model follows a more classical linguistic approach. Both models are built using the Finite-State Transducer toolkit foma. The resultant architecture shows differences in size and structural clarity. While the 'Chunking' model is under half the size of the full de-composed counterpart, the decomposition displays higher structural order. In this paper, we describe the challenges encountered when modelling a language exhibiting distributed exponence and present the first morphological analyser for Nen, with an overall accuracy of 80.3%.

Embeddings of Label Components for Sequence Labeling: A Case Study of Fine-grained Named Entity Recognition [Website][PDF]
Takuma Kato, Kaori Abe, Hiroki Ouchi, Shumpei Miyawaki, Jun Suzuki, and Kentaro Inui 8:00–9:00

In general, the labels used in sequence labeling consist of different types of elements. For example, IOB-format entity labels, such as B-Person and I-Person, can be decomposed into span (B and I) and type information (Person). However, while most sequence labeling models do not consider such label components, the shared components across labels, such as Person, can be beneficial for label prediction. In this work, we propose to integrate label component information as embeddings into models. Through experiments on English and Japanese fine-grained named entity recognition, we demonstrate that the proposed method improves performance, especially for instances with low-frequency labels.

Session 7A Syntax: Tagging, Chunking and Parsing-2

[TACL] A Graph-based Model for Joint Chinese Word Segmentation and Dependency Parsing [Website][PDF]
Hang Yan, Xipeng Qiu, and Xuanjing Huang 8:00–9:00

Chinese word segmentation and dependency parsing are two fundamental tasks for Chinese natural language processing. The dependency parsing is defined on word-level. Therefore, word segmentation is the precondition of dependency parsing, which makes dependency parsing suffer from error propagation and unable to directly make use of the character-level pre-trained language model (such as BERT). In this paper, we propose a graph-based model to integrate Chinese word segmentation and dependency parsing. Different from previous transition-based joint models, our proposed model is more concise, which results in fewer efforts of feature engineering. Our graph-based joint model achieves better performance than previous joint models and state-of-the-art results in both Chinese word segmentation and dependency parsing. Besides, when BERT is combined, our model can substantially reduce the performance gap of dependency parsing between joint models and gold-segmented word-based models. Our code is publicly available at https://github.com/fastnlp/JointCwsParser.

[CL] Abstract Syntax as Interlingua: Scaling Up the Grammatical Framework from Controlled Languages to Robust Pipelines [Website][PDF]
Aarne Ranta, Krasimir Angelov, Normunds Gruzitis, and Prasanth Kolachina 8:00–9:00

Abstract syntax is an interlingual representation used in compilers. Grammatical Framework (GF) applies the abstract syntax idea to natural languages. The development of GF started in 1998, first as a tool for controlled language implementations, where it has gained an established position in both academic and commercial projects. GF provides grammar resources for over 40 languages, enabling accurate generation and translation, as well as grammar engineering tools and components for mobile and Web applications. On the research side, the focus in the last ten years has been on scaling up GF to wide-coverage language processing. The concept of abstract syntax offers a unified view on many other approaches: Universal Dependencies, WordNets, FrameNets, Construction Grammars, and Abstract Meaning Representations. This makes it possible for GF to utilize data from the other approaches and to build robust pipelines. In return, GF can contribute to data-driven approaches by methods to transfer resources from one language to others, to augment data by rule-based generation, to check the consistency of hand-annotated corpora, and to pipe analyses into high-precision semantic back ends. This article gives an overview of the use of abstract syntax as interlingua through both established and emerging NLP applications involving GF.

An Empirical Comparison of Unsupervised Constituency Parsing Methods [Website][PDF]
Jun Li, Yifan Cao, Jiong Cai, Yong Jiang, and Kewei Tu 8:00–9:00

Unsupervised constituency parsing aims to learn a constituency parser from a training corpus without parse tree annotations. While many methods have been proposed to tackle the problem, including statistical and neural methods, their experimental results are often not directly comparable due to discrepancies in datasets, data preprocessing, lexicalization, and evaluation metrics. In this paper, we first examine experimental settings used in previous work and propose to standardize the settings for better comparability between methods. We then empirically compare several existing methods, including decade-old and newly proposed ones, under the standardized settings on English and Japanese, two languages with different branching tendencies. We find that recent models do not show a clear advantage over decade-old models in our experiments. We hope our work can provide new insights into existing methods and facilitate future empirical evaluation of unsupervised constituency parsing.

Do Neural Language Models Show Preferences for Syntactic Formalisms? [Website][PDF]
Artur Kulmizev, Vinit Ravishankar, Mostafa Abdou, and Joakim Nivre 8:00–9:00

Recent work on the interpretability of deep neural language models has concluded that many properties of natural language syntax are encoded in their representational spaces. However, such studies often suffer from limited scope by focusing on a single language and a single linguistic formalism. In this study, we aim to investigate the extent to which the semblance of syntactic structure captured by language models adheres to a surface-syntactic or deep syntactic style of analysis, and whether the patterns are consistent across different languages. We apply a probe for extracting directed dependency trees to BERT and ELMo models trained on 13 different languages, probing for two different syntactic annotation styles: Universal Dependencies (UD), prioritizing deep syntactic relations, and Surface-Syntactic Universal Dependencies (SUD), focusing on surface structure. We find that both models exhibit a preference for UD over SUD — with interesting variations across languages and layers — and that the strength of this preference is correlated with differences in tree shape.

Efficient Constituency Parsing by Pointing [Website][PDF]
Thanh-Tung Nguyen, Xuan-Phi Nguyen, Shafiq Joty, and Xiaoli Li 8:00–9:00

We propose a novel constituency parsing model that casts the parsing problem into a series of pointing tasks. Specifically, our model estimates the likelihood of a span being a legitimate tree constituent via the pointing score corresponding to the boundary words of the span. Our parsing model supports efficient top-down decoding and our learning objective is able to enforce structural consistency without resorting to the expensive CKY inference. The experiments on the standard English Penn Treebank parsing task show that our method achieves 92.78 F1 without using pre-trained models, which is higher than all the existing methods with similar time complexity. Using pre-trained BERT, our model achieves 95.48 F1, which is competitive with the state-of-the-art while being faster. Our approach also establishes new state-of-the-art in Basque and Swedish in the SPMRL shared tasks on multilingual constituency parsing.

Efficient Second-Order TreeCRF for Neural Dependency Parsing [Website][PDF]
Yu Zhang, Zhenghua Li, and Min Zhang 8:00–9:00

In the deep learning (DL) era, parsing models are extremely simplified with little hurt on performance, thanks to the remarkable capability of multi-layer BiLSTMs in context representation. As the most popular graph-based dependency parser due to its high efficiency and performance, the biaffine parser directly scores single dependencies under the arc-factorization assumption, and adopts a very simple local token-wise cross-entropy training loss. This paper for the first time presents a second-order TreeCRF extension to the biaffine parser. For a long time, the complexity and inefficiency of the inside-outside algorithm hinder the popularity of TreeCRF. To address this issue, we propose an effective way to batchify the inside and Viterbi algorithms for direct large matrix operation on GPUs, and to avoid the complex outside algorithm via efficient back-propagation. Experiments and analysis on 27 datasets from 13 languages clearly show that techniques developed before the DL era, such as structural learning (global TreeCRF loss) and high-order modeling are still useful, and can further boost parsing performance over the state-of-the-art biaffine parser, especially for partially annotated training data. We release our code at https://github.com/yzhangcs/crfpar.

Enriched In-Order Linearization for Faster Sequence-to-Sequence Constituent Parsing [Website][PDF]
Daniel Fernández-González and Carlos Gómez-Rodríguez 8:00–9:00

Sequence-to-sequence constituent parsing requires a linearization to represent trees as sequences. Top-down tree linearizations, which can be based on brackets or shift-reduce actions, have achieved the best accuracy to date. In this paper, we show that these results can be improved by using an in-order linearization instead. Based on this observation, we implement an enriched in-order shift-reduce linearization inspired by Vinyals et al. (2015)'s approach, achieving the best accuracy to date on the English PTB dataset among fully-supervised single-model sequence-to-sequence constituent parsers. Finally, we apply deterministic attention mechanisms to match the speed of state-of-the-art transition-based parsers, thus showing that sequence-to-sequence models can match them, not only in accuracy, but also in speed.

Exact yet Efficient Graph Parsing, Bi-directional Locality and the Constructivist Hypothesis [Website][PDF]
Yajie Ye and Weiwei Sun 8:00–9:00

A key problem in processing graph-based meaning representations is graph parsing, i.e. computing all possible derivations of a given graph according to a (competence) grammar. We demonstrate, for the first time, that exact graph parsing can be efficient for large graphs and with large Hyperedge Replacement Grammars (HRGs). The advance is achieved by exploiting locality as terminal edge-adjacency in HRG rules. In particular, we highlight the importance of 1) a terminal edge-first parsing strategy, 2) a categorization of a subclass of HRG, i.e. what we call Weakly Regular Graph Grammar, and 3) distributing argument-structures to both lexical and phrasal rules.

Max-Margin Incremental CCG Parsing [Website][PDF]
Miloš Stanojević and Mark Steedman 8:00–9:00

Incremental syntactic parsing has been an active research area both for cognitive scientists trying to model human sentence processing and for NLP researchers attempting to combine incremental parsing with language modelling for ASR and MT. Most effort has been directed at designing the right transition mechanism, but less has been done to answer the question of what a probabilistic model for those transition parsers should look like. A very incremental transition mechanism of a recently proposed CCG parser when trained in straightforward locally normalised discriminative fashion produces very bad results on English CCGbank. We identify three biases as the causes of this problem: label bias, exposure bias and imbalanced probabilities bias. While known techniques for tackling these biases improve results, they still do not make the parser state of the art. Instead, we tackle all of these three biases at the same time using an improved version of beam search optimisation that minimises all beam search violations instead of minimising only the biggest violation. The new incremental parser gives better results than all previously published incremental CCG parsers, and outperforms even some widely used non-incremental CCG parsers.

Neural Reranking for Dependency Parsing: An Evaluation [Website][PDF]
Bich-Ngoc Do and Ines Rehbein 8:00–9:00

Recent work has shown that neural rerankers can improve results for dependency parsing over the top k trees produced by a base parser. However, all neural rerankers so far have been evaluated on English and Chinese only, both languages with a configurational word order and poor morphology. In the paper, we re-assess the potential of successful neural reranking models from the literature on English and on two morphologically rich(er) languages, German and Czech. In addition, we introduce a new variation of a discriminative reranker based on graph convolutional networks (GCNs). We show that the GCN not only outperforms previous models on English but is the only model that is able to improve results over the baselines on German and Czech. We explain the differences in reranking performance based on an analysis of a) the gold tree ratio and b) the variety in the k-best lists.

Representations of Syntax [MASK] Useful: Effects of Constituency and Dependency Structure in Recursive LSTMs [Website][PDF]
Michael Lepori, Tal Linzen, and R. Thomas McCoy 8:00–9:00

Sequence-based neural networks show significant sensitivity to syntactic structure, but they still perform less well on syntactic tasks than tree-based networks. Such tree-based networks can be provided with a constituency parse, a dependency parse, or both. We evaluate which of these two representational schemes more effectively introduces biases for syntactic structure that increase performance on the subject-verb agreement prediction task. We find that a constituency-based network generalizes more robustly than a dependency-based one, and that combining the two

types of structure does not yield further improvement. Finally, we show that the syntactic robustness of sequential models can be substantially improved by fine-tuning on a small amount of constructed data, suggesting that data augmentation is a viable alternative to explicit constituency structure for imparting the syntactic biases that sequential models are lacking.

Structure-Level Knowledge Distillation For Multilingual Sequence Labeling [Website][PDF]
Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Fei Huang, and Kewei Tu 8:00–9:00

Multilingual sequence labeling is a task of predicting label sequences using a single unified model for multiple languages. Compared with relying on multiple monolingual models, using a multilingual model has the benefit of a smaller model size, easier in online serving, and generalizability to low-resource languages. However, current multilingual models still underperform individual monolingual models significantly due to model capacity limitations. In this paper, we propose to reduce the gap between monolingual models and the unified multilingual model by distilling the structural knowledge of several monolingual models (teachers) to the unified multilingual model (student). We propose two novel KD methods based on structure-level information: (1) approximately minimizes the distance between the student's and the teachers' structure-level probability distributions, (2) aggregates the structure-level knowledge to local distributions and minimizes the distance between two local probability distributions. Our experiments on 4 multilingual tasks with 25 datasets show that our approaches outperform several strong baselines and have stronger zero-shot generalizability than both the baseline model and teacher models.

Demo Session 2B

Time: *8:45–9:30*

LinggleWrite: a Coaching System for Essay Writing [Website] [PDF]
Chung-Ting Tsai, Jhih-Jie Chen, Ching-Yu Yang, and Jason S. Chang

This paper presents LinggleWrite, a writing coach that provides writing suggestions, assesses writing proficiency levels, detects grammatical errors, and offers corrective feedback in response to user's essay. The method involves extracting grammar patterns, training models for automated essay scoring (AES) and grammatical error detection (GED), and finally retrieving plausible corrections from a n-gram search engine. Experiments on public test sets indicate that both AES and GED models achieve state-of-the-art performance. These results show that LinggleWrite is potentially useful in helping learners improve their writing skills.

Session 7B Overview – Tuesday, July 7, 2020 9:00–10:00

Track A *Ethics and NLP-2* Abstracts	Demographics Should Not Be the Reason of Toxicity: Mitigating Discrimination in Text Classifications with Instance Weighting *Zhang, Bai, Zhang, Bai, Zhu, and Zhao* [Website][PDF]	Give Me Convenience and Give Her Death: Who Should Decide What Uses of NLP are Appropriate, and on What Basis? *, Lau, and Balduwin* [Website][PDF]	Is Your Classifier Actually Biased? Measuring Fairness under Uncertainty with Bernstein Bounds *Ethayarajh* [Website][PDF]	Towards Understanding Gender Bias in Relation Extraction *Gaut, Sun, Tang, Huang, Qian, ElSherief, Zhao, Mirza, Belding, Chang, and Wang* [Website][PDF]	
Track B *Interpretability and Analysis of Models for NLP-2* Abstracts	An Analysis of the Utility of Explicit Negative Examples to Improve the Syntactic Abilities of Neural Language Models *Noji and Takamura* [Website][PDF]	Analyzing analytical methods: The case of phonology in neural models of spoken language *Chrupała, Higy, and Alishahi* [Website][PDF]	Make Up Your Mind! Adversarial Generation of Inconsistent Natural Language Explanations *Camburu, Shillingford, Minervini, Lukasiewicz, and Blunsom* [Website][PDF]	Perturbed Masking: Parameter-free Probing for Analyzing and Interpreting BERT *Wu, Chen, Kao, and Liu* [Website][PDF]	Probing for Referential Information in Language Models *Sorodoc, Gulordava, and Boleda* [Website][PDF]
	Quantifying Attention Flow in Transformers *Abnar and Zuidema* [Website][PDF]	Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? *Jacovi and Goldberg* [Website][PDF]	Towards Transparent and Explainable Attention Models *Mohankumar, Nema, Narasimhan, Khapra, Srinivasan, and Ravindran* [Website][PDF]		
Track C *Machine Learning for NLP-8* Abstracts	A Relational Memory-based Embedding Model for Triple Classification and Search Personalization *Nguyen, Nguyen, and Phung* [Website][PDF]	Do you have the right scissors? Tailoring Pre-trained Language Models via Monte-Carlo Methods *Miao, Song, Zhou, and Li* [Website][PDF]	Enhancing Pre-trained Chinese Character Representation with Word-aligned Attention *Li, Yu, Mengge, and Liu* [Website][PDF]	Estimating the influence of auxiliary tasks for multi-task learning of sequence tagging tasks *Schröder and Biemann* [Website][PDF]	Tchebycheff Procedure for Multi-task Text Classification *Mao, Yun, Liu, and Du* [Website][PDF]
Track D *Machine Translation-10* Abstracts	A Graph-based Coarse-to-fine Method for Unsupervised Bilingual Lexicon Induction *Ren, Liu, Zhou, and Ma* [Website][PDF]	A Reinforced Generation of Adversarial Examples for Neural Machine Translation *, Huang, Xie, Dai, and CHEN* [Website][PDF]	A Retrieve-and-Rewrite Initialization Method for Unsupervised Machine Translation *Ren, Wu, Liu, Zhou, and Ma* [Website][PDF]	A Simple and Effective Unified Encoder for Document-Level Machine Translation *Ma, Zhang, and Zhou* [Website][PDF]	Does Multi-Encoder Help? A Case Study on Context-Aware Neural Machine Translation *Li, Liu, Wang, Jiang, Xiao, Zhu, Liu, and* [Website][PDF]
	Knowledge Distillation for Multilingual Unsupervised Neural Machine Translation *Sun, Wang, Chen, Utiyama, Sumita, and Zhao* [Website][PDF]	Learning to Recover from Multi-Modality Errors for Non-Autoregressive Neural Machine Translation *Ran, Lin, Li, and Zhou* [Website][PDF]	Lexically Constrained Neural Machine Translation with Levenshtein Transformer *Susanto, Chollampatt, and Tan* [Website][PDF]	Modeling Word Formation in English—German Neural Machine Translation *Weller-Di Marco and Fraser* [Website][PDF]	

Track E *NLP Applications-6* Abstracts	Camouflaged Chinese Spam Content Detection with Semi-supervised Generative Active Learning *Jiang, Gao, Duan, Kang, Sun, Zhang, and Liu* [Website][PDF]	Distinguish Confusing Law Articles for Legal Judgment Prediction *Xu, Wang, Chen, Pan, Wang, and Zhao* [Website][PDF]	Empowering Active Learning to Jointly Optimize System and User Demands *Lee, Meyer, and Gurevych* [Website][PDF]	Encoder-Decoder Models Can Benefit from Pre-trained Masked Language Models in Grammatical Error Correction *Kaneko, Mita, Kiyono, Suzuki, and Inui* [Website][PDF]	Graph Neural News Recommendation with Unsupervised Preference Disentanglement *Hu, Xu, Li, Yang, Shi, Duan, Xie, and Zhou* [Website][PDF]
	HyperCore: Hyperbolic and Co-graph Representation for Automatic ICD Coding *Cao, Chen, Liu, Zhao, Liu, and Chong* [Website][PDF]	Hyperbolic Capsule Networks for Multi-Label Classification *Chen, Huang, Xiao, and Jing* [Website][PDF]	Identifying Principals and Accessories in a Complex Case based on the Comprehension of Fact Description *Hu, Luo, and Chao* [Website][PDF]	Improving Segmentation for Technical Support Problems *Chauhan and Gupta* [Website][PDF]	Joint Modelling of Emotion and Abusive Language Detection *Rajamanickam, Mishra, Yannakoudakis, and Shutova* [Website][PDF]
	MOOCCube: A Large-scale Data Repository for NLP Applications in MOOCs *Yu, Luo, Xiao, Zhong, Wang, Luo, Wang, Hou, Li, Liu, and Tang* [Website][PDF]	Programming in Natural Language with fuSE: Synthesizing Methods from Spoken Utterances Using Deep Natural Language Understanding *Weigelt, Steurer, Hey, and Tichy* [Website][PDF]	Towards Interpretable Clinical Diagnosis with Bayesian Network Ensembles Stacked on Entity-Aware CNNs *Chen, Dai, Yuan, Lu, and Huang* [Website][PDF]	Toxicity Detection: Does Context Really Matter? *Pavlopoulos, Sorensen, Dixon, Thain, and Androutsopoulos* [Website][PDF]	
Track F *Sentence Level-3* Abstracts	AMR Parsing with Latent Structural Information *Zhou, Zhang, Ji, and Tang* [Website][PDF]	TaPas: Weakly Supervised Table Parsing via Pre-training *Herzig, Nowak, Müller, Piccinno, and Eisenschlos* [Website][PDF]			
Track G *Sentiment Analysis, Stylistic Analysis, and Argument Mining-4* Abstracts	Embarrassingly Simple Unsupervised Aspect Extraction *Tulkens and Cranenburgh* [Website][PDF]	Relation-Aware Collaborative Learning for Unified Aspect-Based Sentiment Analysis *Chen and Qian* [Website][PDF]	Syntax-Aware Opinion Role Labeling with Dependency Graph Convolutional Networks *Zhang, Zhang, Wang, Li, and Zhang* [Website][PDF]	Target Inference in Argument Conclusion Generation *Alshomary, Syed, Potthast, and Wachsmuth* [Website][PDF]	Transition-based Directed Graph Construction for Emotion-Cause Pair Extraction *Fan, Yuan, Du, Gui, Yang, and Xu* [Website][PDF]
Track H *Speech and Multimodality-4* Abstracts	CH-SIMS: A Chinese Multimodal Sentiment Analysis Dataset with Fine-grained Annotation of Modality *Yu, Xu, Meng, Zhu, Ma, Wu, Zou, and Yang* [Website][PDF]	Curriculum Pre-training for End-to-End Speech Translation *Wang, Wu, Liu, Zhou, and Yang* [Website][PDF]	Improving Disfluency Detection by Self-Training a Self-Attentive Model *Jamshid Lou and Johnson* [Website][PDF]	Multimodal Transformer for Multimodal Machine Translation *Yao and Wan* [Website][PDF]	Reasoning with Multimodal Sarcastic Tweets via Modeling Cross-Modality Contrast and Semantic Association *Xu, Zeng, and Mao* [Website][PDF]

	Sentiment and Emotion help Sarcasm? A Multi-task Learning Framework for Multi-Modal Sarcasm, Sentiment and Emotion Analysis *Chauhan, S R, Ekbal, and Bhattacharyya* [Website][PDF]	Towards Emotion-aided Multi-modal Dialogue Act Classification *Saha, Patra, Saha, and Bhattacharyya* [Website][PDF]		
Track I *Student Research Workshop* Abstracts	Building a Japanese Typo Dataset from Wikipedia's Revision History *Tanaka, Murawaki, Kawahara, and Kurohashi* [Website][PDF]	How much complexity does an RNN architecture need to learn syntax-sensitive dependencies? *Bhatt, Bansal, Singh, and Agarwal* [Website][PDF]	Unsupervised Multilingual Sentence Embeddings for Parallel Corpus Mining *Kvapilíková, Artetxe, Labaka, Agirre, and Bojar* [Website][PDF]	

Session 7B Details

Session 7B: Ethics and NLP-2

Demographics Should Not Be the Reason of Toxicity: Mitigating Discrimination in Text Classifications with Instance Weighting [Website][PDF]
Guanhua Zhang, Bing Bai, Junqi Zhang, Kun Bai, Conghui Zhu, and Tiejun Zhao 9:00–10:00

With the recent proliferation of the use of text classifications, researchers have found that there are certain unintended biases in text classification datasets. For example, texts containing some demographic identity-terms (e.g., "gay", "black") are more likely to be abusive in existing abusive language detection datasets. As a result, models trained with these datasets may consider sentences like "She makes me happy to be gay" as abusive simply because of the word "gay." In this paper, we formalize the unintended biases in text classification datasets as a kind of selection bias from the non-discrimination distribution to the discrimination distribution. Based on this formalization, we further propose a model-agnostic debiasing training framework by recovering the non-discrimination distribution using instance weighting, which does not require any extra resources or annotations apart from a pre-defined set of demographic identity-terms. Experiments demonstrate that our method can effectively alleviate the impacts of the unintended biases without significantly hurting models' generalization ability.

Give Me Convenience and Give Her Death: Who Should Decide What Uses of NLP are Appropriate, and on What Basis? [Website][PDF]
kobi leins kobi, Jey Han Lau, and Timothy Baldwin 9:00–10:00

As part of growing NLP capabilities, coupled with an awareness of the ethical dimensions of research, questions have been raised about whether particular datasets and tasks should be deemed off-limits for NLP research. We examine this question with respect to a paper on automatic legal sentencing from EMNLP 2019 which was a source of some debate, in asking whether the paper should have been allowed to be published, who should have been charged with making such a decision, and on what basis. We focus in particular on the role of data statements in ethically assessing research, but also discuss the topic of dual use, and examine the outcomes of similar debates in other scientific disciplines.

Is Your Classifier Actually Biased? Measuring Fairness under Uncertainty with Bernstein Bounds [Website][PDF]
Kawin Ethayarajh 9:00–10:00

Most NLP datasets are not annotated with protected attributes such as gender, making it difficult to measure classification bias using standard measures of fairness (e.g., equal opportunity). However, manually annotating a large dataset with a protected attribute is slow and expensive. Instead of annotating all the examples, can we annotate a subset of them and use that sample to estimate the bias? While it is possible to do so, the smaller this annotated sample is, the less certain we are that the estimate is close to the true bias. In this work, we propose using Bernstein bounds to represent this uncertainty about the bias estimate as a confidence interval. We provide empirical evidence that a 95% confidence interval derived this way consistently bounds the true bias. In quantifying this uncertainty, our method, which we call Bernstein-bounded unfairness, helps prevent classifiers from being deemed biased or unbiased when there is insufficient evidence to make either claim. Our findings suggest that the datasets currently used to measure specific biases are too small to conclusively identify bias except in the most egregious cases. For example, consider a co-reference resolution system that is 5% more accurate on gender-stereotypical sentences – to claim it is biased with 95% confidence, we need a bias-specific dataset that is 3.8 times larger than WinoBias, the largest available.

Towards Understanding Gender Bias in Relation Extraction [Website][PDF]
Andrew Gaut, Tony Sun, Shirlyn Tang, Yuxin Huang, Jing Qian, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang 9:00–10:00

Recent developments in Neural Relation Extraction (NRE) have made significant strides towards Automated Knowledge Base Construction. While much attention has been dedicated towards improvements in accuracy, there have been no attempts in the literature to evaluate social biases exhibited in NRE systems. In this paper, we create WikiGenderBias, a distantly supervised dataset composed of over 45,000 sentences including a 10% human annotated test set for the purpose of analyzing gender bias in relation extraction systems. We find that when extracting spouse-of and hypernym (i.e., occupation) relations, an NRE system performs differently when the gender of the target entity is different. However, such disparity does not appear when extracting relations such as birthDate or birthPlace. We also analyze how existing bias mitigation techniques, such as name anonymization, word embedding debiasing, and data augmentation affect the NRE system in terms of maintaining the test performance and reducing biases. Unfortunately, due to NRE models rely heavily on surface level cues, we find that existing bias mitigation approaches have a negative effect on NRE. Our analysis lays groundwork for future quantifying and mitigating bias in NRE.

Session 7B: Interpretability and Analysis of Models for NLP-2

An Analysis of the Utility of Explicit Negative Examples to Improve the Syntactic Abilities of Neural Language Models [Website][PDF]
Hiroshi Noji and Hiroya Takamura 9:00–10:00

We explore the utilities of explicit negative examples in training neural language models. Negative examples here are incorrect words in a sentence, such as *barks* in **The dogs barks.* Neural language models are commonly trained only on positive examples, a set of sentences in the training data, but recent studies suggest that the models trained in this way are not capable of robustly handling complex syntactic constructions, such as long-distance agreement. In this paper, we first demonstrate that appropriately using negative examples about particular constructions (e.g., subject-verb agreement) will boost the model's robustness on them in English, with a negligible loss of perplexity. The key to our success is an additional margin loss between the log-likelihoods of a correct word and an incorrect word. We then provide a detailed analysis of the trained models. One of our findings is the difficulty of object-relative clauses for RNNs. We find that even with our direct learning signals the models still suffer from resolving agreement across an object-relative clause. Augmentation of training sentences involving the constructions somewhat helps, but the accuracy still does not reach the level of subject-relative clauses. Although not directly cognitively appealing, our method can be a tool to analyze the true architectural limitation of neural models on challenging linguistic constructions.

Analyzing analytical methods: The case of phonology in neural models of spoken language [Website][PDF]
Grzegorz Chrupała, Bertrand Higy, and Afra Alishahi 9:00–10:00

Given the fast development of analysis techniques for NLP and speech processing systems, few systematic studies have been conducted to compare the strengths and weaknesses of each method. As a step in this direction we study the case of representations of phonology in neural network models of spoken language. We use two commonly applied analytical techniques, diagnostic classifiers and representational similarity analysis, to quantify to what extent neural activation patterns encode phonemes and phoneme sequences. We manipulate two factors that can affect the outcome of analysis. First, we investigate the role of learning by comparing neural activations extracted from trained versus randomly-initialized models. Second, we examine the temporal scope of the activations by probing both local activations corresponding to a few milliseconds of the speech signal, and global activations pooled over the whole utterance. We conclude that reporting analysis results with randomly initialized models is crucial, and that global-scope methods tend to yield more consistent and interpretable results and we recommend their use as a complement to local-scope diagnostic methods.

Make Up Your Mind! Adversarial Generation of Inconsistent Natural Language Explanations [Website][PDF]
Oana-Maria Camburu, Brendan Shillingford, Pasquale Minervini, Thomas Lukasiewicz, and Phil Blunsom 9:00–10:00

To increase trust in artificial intelligence systems, a promising research direction consists of designing neural models capable of generating natural language explanations for their predictions. In this work, we show that such models are nonetheless prone to generating mutually inconsistent explanations, such as "Because there is a dog in the image." and "Because there is no dog in the [same] image.", exposing flaws in either the decision-making process of the model or in the generation of the explanations. We introduce a simple yet effective adversarial framework for sanity checking models against the generation of inconsistent natural language explanations. Moreover, as part of the framework, we address the problem of adversarial attacks with full target sequences, a scenario that was not previously addressed in sequence-to-sequence attacks. Finally, we apply our framework on a state-of-the-art neural natural language inference model that provides natural language explanations for its predictions. Our framework shows that this model is capable of generating a significant number of inconsistent explanations.

Perturbed Masking: Parameter-free Probing for Analyzing and Interpreting BERT [Website][PDF]
Zhiyong Wu, Yun Chen, Ben Kao, and Qun Liu 9:00–10:00

By introducing a small set of additional parameters, a *probe* learns to solve specific linguistic tasks (e.g., dependency parsing) in a supervised manner using feature representations (e.g., contextualized embeddings). The effectiveness of such *probing* tasks is taken as evidence that the pre-trained model encodes linguistic knowledge. However, this approach of evaluating a language model is undermined by the uncertainty of the amount of knowledge that is learned by the probe itself. Complementary to those works, we propose a parameter-free probing technique for analyzing pre-trained language models (e.g., BERT). Our method does not require direct supervision from the probing tasks, nor do we introduce additional parameters to the probing process. Our experiments on BERT show that syntactic trees recovered from BERT using our method are significantly better than linguistically-uninformed baselines. We further feed the empirically induced dependency structures into a downstream sentiment classification task and find its improvement compatible with or even superior to a human-designed dependency schema.

Probing for Referential Information in Language Models [Website][PDF]
Ionut-Teodor Sorodoc, Kristina Gulordava, and Gemma Boleda 9:00–10:00

Language models keep track of complex information about the preceding context – including, e.g., syntactic relations in a sentence. We investigate whether they also capture information beneficial for resolving pronominal anaphora in English. We analyze two state of the art models with LSTM and Transformer architectures, via probe tasks and analysis on a coreference annotated corpus. The Transformer outperforms the LSTM in all analyses. Our results suggest that language models are more successful at learning grammatical constraints than they are at learning truly referential

information, in the sense of capturing the fact that we use language to refer to entities in the world. However, we find traces of the latter aspect, too.

Quantifying Attention Flow in Transformers [Website][PDF]
Samira Abnar and Willem Zuidema 9:00–10:00

In the Transformer model, "self-attention" combines information from attended embeddings into the representation of the focal embedding in the next layer. Thus, across layers of the Transformer, information originating from different tokens gets increasingly mixed. This makes attention weights unreliable as explanations probes. In this paper, we consider the problem of quantifying this flow of information through self-attention. We propose two methods for approximating the attention to input tokens given attention weights, attention rollout and attention flow, as post hoc methods when we use attention weights as the relative relevance of the input tokens. We show that these methods give complementary views on the flow of information, and compared to raw attention, both yield higher correlations with importance scores of input tokens obtained using an ablation method and input gradients.

Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? [Website][PDF]
Alon Jacovi and Yoav Goldberg 9:00–10:00

With the growing popularity of deep-learning based NLP models, comes a need for interpretable systems. But what is interpretability, and what constitutes a high-quality interpretation? In this opinion piece we reflect on the current state of interpretability evaluation research. We call for more clearly differentiating between different desired criteria an interpretation should satisfy, and focus on the faithfulness criteria. We survey the literature with respect to faithfulness evaluation, and arrange the current approaches around three assumptions, providing an explicit form to how faithfulness is "defined" by the community. We provide concrete guidelines on how evaluation of interpretation methods should and should not be conducted. Finally, we claim that the current binary definition for faithfulness sets a potentially unrealistic bar for being considered faithful. We call for discarding the binary notion of faithfulness in favor of a more graded one, which we believe will be of greater practical utility.

Towards Transparent and Explainable Attention Models [Website][PDF]
Akash Kumar Mohankumar, Preksha Nema, Sharan Narasimhan, Mitesh M. Khapra, Balaji Vasan Srinivasan, and Balaraman Ravindran 9:00–10:00

Recent studies on interpretability of attention distributions have led to notions of faithful and plausible explanations for a model's predictions. Attention distributions can be considered a faithful explanation if a higher attention weight implies a greater impact on the model's prediction. They can be considered a plausible explanation if they provide a human-understandable justification for the model's predictions. In this work, we first explain why current attention mechanisms in LSTM based encoders can neither provide a faithful nor a plausible explanation of the model's predictions. We observe that in LSTM based encoders the hidden representations at different time-steps are very similar to each other (high conicity) and attention weights in these situations do not carry much meaning because even a random permutation of the attention weights does not affect the model's predictions. Based on experiments on a wide variety of tasks and datasets, we observe attention distributions often attribute the model's predictions to unimportant words such as punctuation and fail to offer a plausible explanation for the predictions. To make attention mechanisms more faithful and plausible, we propose a modified LSTM cell with a diversity-driven training objective that ensures that the hidden representations learned at different time steps are diverse. We show that the resulting attention distributions offer more transparency as they (i) provide a more precise importance ranking of the hidden states (ii) are better indicative of words important for the model's predictions (iii) correlate better with gradient-based attribution methods. Human evaluations indicate that the attention distributions learned by our model offer a plausible explanation of the model's predictions. Our code has been made publicly available at https://github.com/akashkm99/Interpretable-Attention

Session 7B: Machine Learning for NLP-8

A Relational Memory-based Embedding Model for Triple Classification and Search Personalization [Website][PDF]
Dai Quoc Nguyen, Tu Nguyen, and Dinh Phung 9:00–10:00

Knowledge graph embedding methods often suffer from a limitation of memorizing valid triples to predict new ones for triple classification and search personalization problems. To this end, we introduce a novel embedding model, named R-MeN, that explores a relational memory network to encode potential dependencies in relationship triples. R-MeN considers each triple as a sequence of 3 input vectors that recurrently interact with a memory using a transformer self-attention mechanism. Thus R-MeN encodes new information from interactions between the memory and each input vector to return a corresponding vector. Consequently, R-MeN feeds these 3 returned vectors to a convolutional neural network-based decoder to produce a scalar score for the triple. Experimental results show that our proposed R-MeN obtains state-of-the-art results on SEARCH17 for the search personalization task, and on WN11 and FB13 for the triple classification task.

Do you have the right scissors? Tailoring Pre-trained Language Models via Monte-Carlo Methods [Website][PDF]
Ning Miao, Yuxuan Song, Hao Zhou, and Lei Li 9:00–10:00

It has been a common approach to pre-train a language model on a large corpus and fine-tune it on task-specific data. In practice, we observe that fine-tuning a pre-trained model on a small dataset may lead to over- and/or under-estimate problem. In this paper, we propose MC-Tailor, a novel method to alleviate the above issue in text generation tasks by truncating and transferring the probability mass from over-estimated regions to under-estimated ones. Experiments on a variety of text generation datasets show that MC-Tailor consistently and significantly outperforms the fine-tuning approach.

Enhancing Pre-trained Chinese Character Representation with Word-aligned Attention [Website][PDF]
Yanzeng Li, Bowen Yu, Xue Mengge, and Tingwen Liu 9:00–10:00

Most Chinese pre-trained models take character as the basic unit and learn representation according to character's external contexts, ignoring the semantics expressed in the word, which is the smallest meaningful utterance in Chinese. Hence, we propose a novel word-aligned attention to exploit explicit word information, which is complementary to various character-based Chinese pre-trained language models. Specifically, we devise a pooling mechanism to align the character-level attention to the word level and propose to alleviate the potential issue of segmentation error propagation by multi-source information fusion. As a result, word and character information are explicitly integrated at the fine-tuning procedure. Experimental results on five Chinese NLP benchmark tasks demonstrate that our method achieves significant improvements against BERT, ERNIE and BERT-wwm.

Estimating the influence of auxiliary tasks for multi-task learning of sequence tagging tasks [Website][PDF]
Fynn Schröder and Chris Biemann 9:00–10:00

Multi-task learning (MTL) and transfer learning (TL) are techniques to overcome the issue of data scarcity when training state-of-the-art neural networks. However, finding beneficial auxiliary datasets for MTL or TL is a time- and resource-consuming trial-and-error approach. We propose new methods to automatically assess the similarity of sequence tagging datasets to identify beneficial auxiliary data for MTL or TL setups. Our methods can compute the similarity between any two sequence tagging datasets, i.e. they do not need to be annotated with the same tagset or multiple labels in parallel. Additionally, our methods take tokens and their labels into account, which is more robust than only using either of them as an information source, as conducted in prior work. We empirically show that our similarity measures correlate with the change in test score of neural networks that use the auxiliary dataset for MTL to increase the main task performance. We provide an efficient, open-source implementation.

Tchebycheff Procedure for Multi-task Text Classification [Website][PDF]
Yuren Mao, Shuang Yun, Weiwei Liu, and Bo Du 9:00–10:00

Multi-task Learning methods have achieved great progress in text classification. However, existing methods assume that multi-task text classification problems are convex multiobjective optimization problems, which is unrealistic in real-world applications. To address this issue, this paper presents a novel Tchebycheff procedure to optimize the multi-task classification problems without convex assumption. The extensive experiments back up our theoretical analysis and validate the superiority of our proposals.

Session 7B: Machine Translation-10

A Graph-based Coarse-to-fine Method for Unsupervised Bilingual Lexicon Induction [Website][PDF]
Shuo Ren, Shujie Liu, Ming Zhou, and Shuai Ma 9:00–10:00

Unsupervised bilingual lexicon induction is the task of inducing word translations from monolingual corpora of two languages. Recent methods are mostly based on unsupervised cross-lingual word embeddings, the key to which is to find initial solutions of word translations, followed by the learning and refinement of mappings between the embedding spaces of two languages. However, previous methods find initial solutions just based on word-level information, which may be (1) limited and inaccurate, and (2) prone to contain some noise introduced by the insufficiently pre-trained embeddings of some words. To deal with those issues, in this paper, we propose a novel graph-based paradigm to induce bilingual lexicons in a coarse-to-fine way. We first build a graph for each language with its vertices representing different words. Then we extract word cliques from the graphs and map the cliques of two languages. Based on that, we induce the initial word translation solution with the central words of the aligned cliques. This coarse-to-fine approach not only leverages clique-level information, which is richer and more accurate, but also effectively reduces the bad effect of the noise in the pre-trained embeddings. Finally, we take the initial solution as the seed to learn cross-lingual embeddings, from which we induce bilingual lexicons. Experiments show that our approach improves the performance of bilingual lexicon induction compared with previous methods.

A Reinforced Generation of Adversarial Examples for Neural Machine Translation [Website][PDF]
wei zou wei, Shujian Huang, Jun Xie, Xinyu Dai, and Jiajun CHEN 9:00–10:00

Neural machine translation systems tend to fail on less decent inputs despite its significant efficacy, which may significantly harm the credibility of these systems—fathoming how and when neural-based systems fail in such cases is critical for industrial maintenance. Instead of collecting and analyzing bad cases using limited handcrafted error features, here we investigate this issue by generating adversarial examples via a new paradigm based on reinforcement learning. Our paradigm could expose pitfalls for a given performance metric, e.g., BLEU, and could target any given neural machine translation architecture. We conduct experiments of adversarial attacks on two mainstream neural machine translation architectures, RNN-search, and Transformer. The results show that our method efficiently produces stable attacks with meaning-preserving adversarial examples. We also present a qualitative and quantitative analysis for the preference pattern of the attack, demonstrating its capability of pitfall exposure.

A Retrieve-and-Rewrite Initialization Method for Unsupervised Machine Translation [Website][PDF]
Shuo Ren, Yu Wu, Shujie Liu, Ming Zhou, and Shuai Ma 9:00–10:00

The commonly used framework for unsupervised machine translation builds initial translation models of both translation directions, and then performs iterative back-translation to jointly boost their translation performance. The initialization stage is very important since bad initialization may wrongly squeeze the search space, and too much noise introduced in this stage may hurt the final performance. In this paper, we propose a novel retrieval and rewriting based method to better initialize unsupervised translation models. We first retrieve semantically comparable sentences from monolingual corpora of two languages and then rewrite the target side to minimize the semantic gap between the source and retrieved targets with a designed rewriting model. The rewritten sentence pairs are used to initialize SMT models which are used to generate pseudo data for two NMT models, followed by the iterative back-translation. Experiments show that our method can build better initial unsupervised translation models and improve the final translation performance by over 4 BLEU scores. Our code is released at https://github.com/Imagist-Shuo/RRforUNMT.git.

A Simple and Effective Unified Encoder for Document-Level Machine Translation [Website][PDF]
Shuming Ma, Dongdong Zhang, and Ming Zhou 9:00–10:00

Most of the existing models for document-level machine translation adopt dual-encoder structures. The representation of the source sentences and the document-level contexts[2] are modeled with two separate encoders. Although these models can make use of the document-level contexts, they do not fully model the interaction between the contexts and the source sentences, and can not directly adapt to the recent pre-training models (e.g., BERT) which encodes multiple sentences with a single encoder. In this work, we propose a simple and effective unified encoder that can outperform the baseline models of dual-encoder models in terms of BLEU and METEOR scores. Moreover, the pre-training models can further boost the performance of our proposed model.

Does Multi-Encoder Help? A Case Study on Context-Aware Neural Machine Translation [Website][PDF]
Bei Li, Hui Liu, Ziyang Wang, Yufan Jiang, Tong Xiao, Jingbo Zhu, Tongran Liu, and changliang li changliang 9:00–10:00

In encoder-decoder neural models, multiple encoders are in general used to represent the contextual information in addition to the individual sentence. In this paper, we investigate multi-encoder approaches in document-level neural machine translation (NMT). Surprisingly, we find that the context encoder does not only encode the surrounding sentences but also behaves as a noise generator. This makes us rethink the real benefits of multi-encoder in context-aware translation - some of the improvements come from robust training. We compare several methods that introduce noise and/or well-tuned dropout setup into the training of these encoders. Experimental results show that noisy training plays an important role in multi-encoder-based NMT, especially when the training data is small. Also, we establish a new state-of-the-art on IWSLT Fr-En task by careful use of noise generation and dropout methods.

[2] In this work, document-level contexts denote the surrounding sentences of the current source sentence.

Knowledge Distillation for Multilingual Unsupervised Neural Machine Translation [Website][PDF]
Haipeng Sun, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao 9:00–10:00

Unsupervised neural machine translation (UNMT) has recently achieved remarkable results for several language pairs. However, it can only translate between a single language pair and cannot produce translation results for multiple language pairs at the same time. That is, research on multilingual UNMT has been limited. In this paper, we empirically introduce a simple method to translate between thirteen languages using a single encoder and a single decoder, making use of multilingual data to improve UNMT for all language pairs. On the basis of the empirical findings, we propose two knowledge distillation methods to further enhance multilingual UNMT performance. Our experiments on a dataset with English translated to and from twelve other languages (including three language families and six language branches) show remarkable results, surpassing strong unsupervised individual baselines while achieving promising performance between non-English language pairs in zero-shot translation scenarios and alleviating poor performance in low-resource language pairs.

Learning to Recover from Multi-Modality Errors for Non-Autoregressive Neural Machine Translation [Website][PDF]
Qiu Ran, Yankai Lin, Peng Li, and Jie Zhou 9:00–10:00

Non-autoregressive neural machine translation (NAT) predicts the entire target sequence simultaneously and significantly accelerates inference process. However, NAT discards the dependency information in a sentence, and thus inevitably suffers from the multi-modality problem: the target tokens may be provided by different possible translations, often causing token repetitions or missing. To alleviate this problem, we propose a novel semi-autoregressive model RecoverSAT in this work, which generates a translation as a sequence of segments. The segments are generated simultaneously while each segment is predicted token-by-token. By dynamically determining segment length and deleting repetitive segments, RecoverSAT is capable of recovering from repetitive and missing token errors. Experimental results on three widely-used benchmark datasets show that our proposed model achieves more than 4 times speedup while maintaining comparable performance compared with the corresponding autoregressive model.

Lexically Constrained Neural Machine Translation with Levenshtein Transformer [Website][PDF]
Raymond Hendy Susanto, Shamil Chollampatt, and Liling Tan 9:00–10:00

This paper proposes a simple and effective algorithm for incorporating lexical constraints in neural machine translation. Previous work either required re-training existing models with the lexical constraints or incorporating them during beam search decoding with significantly higher computational overheads. Leveraging the flexibility and speed of a recently proposed Levenshtein Transformer model (Gu et al., 2019), our method injects terminology constraints at inference time without any impact on decoding speed. Our method does not require any modification to the training procedure and can be easily applied at runtime with custom dictionaries. Experiments on English-German WMT datasets show that our approach improves an unconstrained baseline and previous approaches.

Modeling Word Formation in English—German Neural Machine Translation [Website][PDF]
Marion Weller-Di Marco and Alexander Fraser 9:00–10:00

This paper studies strategies to model word formation in NMT using rich linguistic information, namely a word segmentation approach that goes beyond splitting into substrings by considering fusional morphology. Our linguistically sound segmentation is combined with a method for target-side inflection to accommodate modeling word formation. The best system variants employ source-side morphological analysis and model complex target-side words, improving over a standard system.

Session 7B: NLP Applications-6

Camouflaged Chinese Spam Content Detection with Semi-supervised Generative Active Learning [Website][PDF]
Zhuoren Jiang, Zhe Gao, Yu Duan, Yangyang Kang, Changlong Sun, Qiong Zhang, and Xiaozhong Liu 9:00–10:00

We propose a Semi-supervIsed GeNerative Active Learning (SIGNAL) model to address the imbalance, efficiency, and text camouflage problems of Chinese text spam detection task. A "self-diversity" criterion is proposed for measuring the "worthiness" of a candidate for annotation. A semi-supervised variational autoencoder with masked attention learning approach and a character variation graph-enhanced augmentation procedure are proposed for data augmentation. The preliminary experiment demonstrates the proposed SIGNAL model is not only sensitive to spam sample selection, but also can improve the performance of a series of conventional active learning models for Chinese spam detection task. To the best of our knowledge, this is the first work to integrate active learning and semi-supervised generative learning for text spam detection.

Distinguish Confusing Law Articles for Legal Judgment Prediction [Website][PDF]
Nuo Xu, Pinghui Wang, Long Chen, Li Pan, Xiaoyan Wang, and Junzhou Zhao 9:00–10:00

Legal Judgement Prediction (LJP) is the task of automatically predicting a law case's judgment results given a text describing the case's facts, which has great prospects in judicial assistance systems and handy services for the public. In practice, confusing charges are often presented, because law cases applicable to similar law articles are easily misjudged. To address this issue, existing work relies heavily on domain experts, which hinders its application in different law systems. In this paper, we present an end-to-end model, LADAN, to solve the task of LJP. To distinguish confusing charges, we propose a novel graph neural network, GDL, to automatically learn subtle differences between confusing law articles, and also design a novel attention mechanism that fully exploits the learned differences to attentively extract effective discriminative features from fact descriptions. Experiments conducted on real-world datasets demonstrate the superiority of our LADAN.

Empowering Active Learning to Jointly Optimize System and User Demands [Website][PDF]
Ji-Ung Lee, Christian M. Meyer, and Iryna Gurevych 9:00–10:00

Existing approaches to active learning maximize the system performance by sampling unlabeled instances for annotation that yield the most efficient training. However, when active learning is integrated with an end-user application, this can lead to frustration for participating users, as they spend time labeling instances that they would not otherwise be interested in reading. In this paper, we propose a new active learning approach that jointly optimizes the seemingly counteracting objectives of the active learning system (training efficiently) and the user (receiving useful instances). We study our approach in an educational application, which particularly benefits from this technique as the system needs to rapidly learn to predict the appropriateness of an exercise to a particular user, while the users should receive only exercises that match their skills. We evaluate multiple learning strategies and user types with data from real users and find that our joint approach better satisfies both objectives when alternative methods lead to many unsuitable exercises for end users.

Encoder-Decoder Models Can Benefit from Pre-trained Masked Language Models in Grammatical Error Correction [Website][PDF]
Masahiro Kaneko, Masato Mita, Shun Kiyono, Jun Suzuki, and Kentaro Inui 9:00–10:00

This paper investigates how to effectively incorporate a pre-trained masked language model (MLM), such as BERT, into an encoder-decoder (EncDec) model for grammatical error correction (GEC). The answer to this question is not as straightforward as one might expect because the previous common methods for incorporating a MLM into an EncDec model have potential drawbacks when applied to GEC. For example, the distribution of the inputs to a GEC model can be considerably different (erroneous, clumsy, etc.) from that of the corpora used for pre-training MLMs; however, this issue is not addressed in the previous methods. Our experiments show that our proposed method, where we first fine-tune a MLM with a given GEC corpus and then use the output of the fine-tuned MLM as additional features in the GEC model, maximizes the benefit of the MLM. The best-performing model achieves state-of-the-art performances on the BEA-2019 and CoNLL-2014 benchmarks. Our code is publicly available at: https://github.com/kanekomasahiro/bert-gec.

Graph Neural News Recommendation with Unsupervised Preference Disentanglement [Website][PDF]
Linmei Hu, Siyong Xu, Chen Li, Cheng Yang, Chuan Shi, Nan Duan, Xing Xie, and Ming Zhou 9:00–10:00

With the explosion of news information, personalized news recommendation has become very important for users to quickly find their interested contents. Most existing methods usually learn the representations of users and news from news contents for recommendation. However, they seldom consider high-order connectivity underlying the user-news interactions. Moreover, existing methods failed to disentangle a user's latent preference factors which cause her clicks on different news. In this paper, we model the user-news interactions as a bipartite graph and propose a novel Graph Neural News Recommendation model with Unsupervised Preference Disentanglement, named GNUD. Our model can encode high-order relationships into user and news representations by information propagation along the graph. Furthermore, the learned representations are disentangled with latent preference factors by a neighborhood routing algorithm, which can enhance expressiveness and interpretability. A preference regularizer is also designed to force each disentangled subspace to independently reflect an isolated preference, improving the quality of the disentangled representations. Experimental results on real-world news datasets demonstrate that our proposed model can effectively improve the performance of news recommendation and outperform state-of-the-art news recommendation methods.

HyperCore: Hyperbolic and Co-graph Representation for Automatic ICD Coding [Website][PDF]
Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Shengping Liu, and Weifeng Chong 9:00–10:00

The International Classification of Diseases (ICD) provides a standardized way for classifying diseases, which endows each disease with a unique code. ICD coding aims to assign proper ICD codes to a medical record. Since manual coding is very laborious and prone to errors, many methods have been proposed for the automatic ICD coding task. However, most of existing methods independently predict each code, ignoring two important characteristics: Code Hierarchy and Code Co-occurrence. In this paper, we propose a Hyperbolic and Co-graph Representation method (HyperCore) to address the above problem. Specifically, we propose a hyperbolic representation method to leverage the code hierarchy. Moreover, we propose a graph convolutional network to utilize the code co-occurrence. Experimental results on two widely used datasets demonstrate that our proposed model outperforms previous state-of-the-art methods.

Hyperbolic Capsule Networks for Multi-Label Classification [Website][PDF]
Boli Chen, Xin Huang, Lin Xiao, and Liping Jing 9:00–10:00

Although deep neural networks are effective at extracting high-level features, classification methods usually encode an input into a vector representation via simple feature aggregation operations (e.g. pooling). Such operations limit the performance. For instance, a multi-label document may contain several concepts. In this case, one vector can not sufficiently capture its salient and discriminative content. Thus, we propose Hyperbolic Capsule Networks (HyperCaps) for Multi-Label Classification (MLC), which have two merits. First, hyperbolic capsules are designed to capture fine-grained document information for each label, which has the ability to characterize complicated structures among labels and documents. Second, Hyperbolic Dynamic Routing (HDR) is introduced to aggregate hyperbolic capsules in a label-aware manner, so that the label-level discriminative information can be preserved along the depth of neural networks. To efficiently handle large-scale MLC datasets, we additionally present a new routing method to adaptively adjust the capsule number during routing. Extensive experiments are conducted on four benchmark datasets. Compared with the state-of-the-art methods, HyperCaps significantly improves the performance of MLC especially on tail labels.

Identifying Principals and Accessories in a Complex Case based on the Comprehension of Fact Description [Website][PDF]
Yakun Hu, Zhunchen Luo, and Wenhan Chao 9:00–10:00

In this paper, we study the problem of identifying the principals and accessories from the fact description with multiple defendants in a criminal case. We treat the fact descriptions as narrative texts and the defendants as roles over the narrative story. We propose to model the defendants with *behavioral semantic information* and *statistical characteristics*, then learning the importances of defendants within a learning-to-rank framework. Experimental results on a real-world dataset demonstrate the behavior analysis can effectively model the defendants' impacts in a complex case.

Improving Segmentation for Technical Support Problems [Website][PDF]
Kushal Chauhan and Abhirut Gupta 9:00–10:00

Technical support problems are often long and complex. They typically contain user descriptions of the problem, the setup, and steps for attempted resolution. Often they also contain various non-natural language text elements like outputs of commands, snippets of code, error messages or stack traces. These elements contain potentially crucial information for problem resolution. However, they cannot be correctly parsed by tools designed for natural language. In this paper, we address the problem of segmentation for technical support questions. We formulate the problem as a sequence labelling task, and study the performance of state of the art approaches. We compare this against an intuitive contextual sentence-level classification baseline, and a state of the art supervised text-segmentation approach. We also introduce a novel component of combining contextual embeddings from multiple language models pre-trained on different data sources, which achieves a marked improvement over using embeddings from a single pre-trained language model. Finally, we also demonstrate the usefulness of such segmentation with improvements on the downstream task of answer retrieval.

Joint Modelling of Emotion and Abusive Language Detection [Website][PDF]
Santhosh Rajamanickam, Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova 9:00–10:00

The rise of online communication platforms has been accompanied by some undesirable effects, such as the proliferation of aggressive and abusive behaviour online. Aiming to tackle this problem, the natural language processing (NLP) community has experimented with a range of techniques for abuse detection. While achieving substantial success, these methods have so far only focused on modelling the linguistic properties of the comments and the online communities of users, disregarding the emotional state of the users and how this might affect their language. The latter is, however, inextricably linked to abusive behaviour. In this paper, we present the first joint model of emotion and abusive language detection, experimenting in a multi-task learning framework that allows one task to inform the other. Our results demonstrate that incorporating affective features leads to significant improvements in abuse detection performance across datasets.

MOOCCube: A Large-scale Data Repository for NLP Applications in MOOCs [Website][PDF]
Jifan Yu, Gan Luo, Tong Xiao, Qingyang Zhong, Yuquan Wang, wenzheng feng wenzheng, Junyi Luo, Chenyu Wang, Lei Hou, Juanzi Li, Zhiyuan Liu, and Jie Tang 9:00–10:00

The prosperity of Massive Open Online Courses (MOOCs) provides fodder for many NLP and AI research for education applications, e.g., course concept extraction, prerequisite relation discovery, etc. However, the publicly available datasets of MOOC are limited in size with few types of data, which hinders advanced models and novel attempts in related topics. Therefore, we present MOOCCube, a large-scale data repository of over 700 MOOC courses, 100k con-

cepts, 8 million student behaviors with an external resource. Moreover, we conduct a prerequisite discovery task as an example application to show the potential of MOOCCube in facilitating relevant research. The data repository is now available at http://moocdata.cn/data/MOOCCube.

Programming in Natural Language with fuSE: Synthesizing Methods from Spoken Utterances Using Deep Natural Language Understanding [Website][PDF]
Sebastian Weigelt, Vanessa Steurer, Tobias Hey, and Walter F. Tichy 9:00–10:00

The key to effortless end-user programming is natural language. We examine how to teach intelligent systems new functions, expressed in natural language. As a first step, we collected 3168 samples of teaching efforts in plain English. Then we built fuSE, a novel system that translates English function descriptions into code. Our approach is three-tiered and each task is evaluated separately. We first classify whether an intent to teach new functionality is present in the utterance (accuracy: 97.7% using BERT). Then we analyze the linguistic structure and construct a semantic model (accuracy: 97.6% using a BiLSTM). Finally, we synthesize the signature of the method, map the intermediate steps (instructions in the method body) to API calls and inject control structures (F_1: 67.0% with information retrieval and knowledge-based methods). In an end-to-end evaluation on an unseen dataset fuSE synthesized 84.6% of the method signatures and 79.2% of the API calls correctly.

Towards Interpretable Clinical Diagnosis with Bayesian Network Ensembles Stacked on Entity-Aware CNNs [Website][PDF]
Jun Chen, Xiaoya Dai, Quan Yuan, Chao Lu, and Haifeng Huang 9:00–10:00

The automatic text-based diagnosis remains a challenging task for clinical use because it requires appropriate balance between accuracy and interpretability. In this paper, we attempt to propose a solution by introducing a novel framework that stacks Bayesian Network Ensembles on top of Entity-Aware Convolutional Neural Networks (CNN) towards building an accurate yet interpretable diagnosis system. The proposed framework takes advantage of the high accuracy and generality of deep neural networks as well as the interpretability of Bayesian Networks, which is critical for AI-empowered healthcare. The evaluation conducted on the real Electronic Medical Record (EMR) documents from hospitals and annotated by professional doctors proves that, the proposed framework outperforms the previous automatic diagnosis methods in accuracy performance and the diagnosis explanation of the framework is reasonable.

Toxicity Detection: Does Context Really Matter? [Website][PDF]
John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos 9:00–10:00

Moderation is crucial to promoting healthy online discussions. Although several 'toxicity' detection datasets and models have been published, most of them ignore the context of the posts, implicitly assuming that comments may be judged independently. We investigate this assumption by focusing on two questions: (a) does context affect the human judgement, and (b) does conditioning on context improve performance of toxicity detection systems? We experiment with Wikipedia conversations, limiting the notion of context to the previous post in the thread and the discussion title. We find that context can both amplify or mitigate the perceived toxicity of posts. Moreover, a small but significant subset of manually labeled posts (5% in one of our experiments) end up having the opposite toxicity labels if the annotators are not provided with context. Surprisingly, we also find no evidence that context actually improves the performance of toxicity classifiers, having tried a range of classifiers and mechanisms to make them context aware. This points to the need for larger datasets of comments annotated in context. We make our code and data publicly available.

Session 7B Semantics: Sentence Level-3

AMR Parsing with Latent Structural Information [Website][PDF]
Qiji Zhou, Yue Zhang, Donghong Ji, and Hao Tang 9:00–10:00

Abstract Meaning Representations (AMRs) capture sentence-level semantics structural representations to broad-coverage natural sentences. We investigate parsing AMR with explicit dependency structures and interpretable latent structures. We generate the latent soft structure without additional annotations, and fuse both dependency and latent structure via an extended graph neural networks. The fused structural information helps our experiments results to achieve the best reported results on both AMR 2.0 (77.5% Smatch F1 on LDC2017T10) and AMR 1.0 ((71.8% Smatch F1 on LDC2014T12).

TaPas: Weakly Supervised Table Parsing via Pre-training [Website][PDF]
Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos
9:00–10:00

Answering natural language questions over tables is usually seen as a semantic parsing task. To alleviate the collection cost of full logical forms, one popular approach focuses on weak supervision consisting of denotations instead of logical forms. However, training semantic parsers from weak supervision poses difficulties, and in addition, the generated logical forms are only used as an intermediate step prior to retrieving the denotation. In this paper, we present TaPas, an approach to question answering over tables without generating logical forms. TaPas trains from weak supervision, and predicts the denotation by selecting table cells and optionally applying a corresponding aggregation operator to such selection. TaPas extends BERT's architecture to encode tables as input, initializes from an effective joint pre-training of text segments and tables crawled from Wikipedia, and is trained end-to-end. We experiment with three different semantic parsing datasets, and find that TaPas outperforms or rivals semantic parsing models by improving state-of-the-art accuracy on SQA from 55.1 to 67.2 and performing on par with the state-of-the-art on WikiSQL and WikiTQ, but with a simpler model architecture. We additionally find that transfer learning, which is trivial in our setting, from WikiSQL to WikiTQ, yields 48.7 accuracy, 4.2 points above the state-of-the-art.

Session 7B: Sentiment Analysis, Stylistic Analysis, and Argument Mining-4

Embarrassingly Simple Unsupervised Aspect Extraction [Website][PDF]
Stéphan Tulkens and Andreas van Cranenburgh 9:00–10:00

We present a simple but effective method for aspect identification in sentiment analysis. Our unsupervised method only requires word embeddings and a POS tagger, and is therefore straightforward to apply to new domains and languages. We introduce Contrastive Attention (CAt), a novel single-head attention mechanism based on an RBF kernel, which gives a considerable boost in performance and makes the model interpretable. Previous work relied on syntactic features and complex neural models. We show that given the simplicity of current benchmark datasets for aspect extraction, such complex models are not needed. The code to reproduce the experiments reported in this paper is available at https://github.com/clips/cat.

Relation-Aware Collaborative Learning for Unified Aspect-Based Sentiment Analysis [Website][PDF]
Zhuang Chen and Tieyun Qian 9:00–10:00

Aspect-based sentiment analysis (ABSA) involves three subtasks, i.e., aspect term extraction, opinion term extraction, and aspect-level sentiment classification. Most existing studies focused on one of these subtasks only. Several recent researches made successful attempts to solve the complete ABSA problem with a unified framework. However, the interactive relations among three subtasks are still under-exploited. We argue that such relations encode collaborative signals between different subtasks. For example, when the opinion term is *"delicious"*, the aspect term must be *"food"* rather than *"place"*. In order to fully exploit these relations, we propose a Relation-Aware Collaborative Learning (RACL) framework which allows the subtasks to work coordinately via the multi-task learning and relation propagation mechanisms in a stacked multi-layer network. Extensive experiments on three real-world datasets demonstrate that RACL significantly outperforms the state-of-the-art methods for the complete ABSA task.

Syntax-Aware Opinion Role Labeling with Dependency Graph Convolutional Networks [Website][PDF]
Bo Zhang, Yue Zhang, Rui Wang, Zhenghua Li, and Min Zhang 9:00–10:00

Opinion role labeling (ORL) is a fine-grained opinion analysis task and aims to answer "who expressed what kind of sentiment towards what?". Due to the scarcity of labeled data, ORL remains challenging for data-driven methods. In this work, we try to enhance neural ORL models with syntactic knowledge by comparing and integrating different representations. We also propose dependency graph convolutional networks (DEPGCN) to encode parser information at different processing levels. In order to compensate for parser inaccuracy and reduce error propagation, we introduce multi-task learning (MTL) to train the parser and the ORL model simultaneously. We verify our methods on the benchmark MPQA corpus. The experimental results show that syntactic information is highly valuable for ORL, and our final MTL model effectively boosts the F1 score by 9.29 over the syntax-agnostic baseline. In addition, we find that the contributions from syntactic knowledge do not fully overlap with contextualized word representations (BERT). Our best model achieves 4.34 higher F1 score than the current state-ofthe-art.

Target Inference in Argument Conclusion Generation [Website][PDF]
Milad Alshomary, Shahbaz Syed, Martin Potthast, and Henning Wachsmuth 9:00–10:00

In argumentation, people state premises to reason towards a conclusion. The conclusion conveys a stance towards some target, such as a concept or statement. Often, the conclusion remains implicit, though, since it is self-evident in a discussion or left out for rhetorical reasons. However, the conclusion is key to understanding an argument and, hence, to any application that processes argumentation. We thus study the question to what extent an argument's conclusion can be reconstructed from its premises. In particular, we argue here that a decisive step is to infer a conclusion's target, and we hypothesize that this target is related to the premises' targets. We develop two complementary target inference approaches: one ranks premise targets and selects the top-ranked target as the conclusion target, the other finds a new conclusion target in a learned embedding space using a triplet neural network. Our evaluation on corpora from two domains indicates that a hybrid of both approaches is best, outperforming several strong baselines. According to human annotators, we infer a reasonably adequate conclusion target in 89% of the cases.

Transition-based Directed Graph Construction for Emotion-Cause Pair Extraction [Website][PDF]
Chuang Fan, Chaofa Yuan, Jiachen Du, Lin Gui, Min Yang, and Ruifeng Xu 9:00–10:00

Emotion-cause pair extraction aims to extract all potential pairs of emotions and corresponding causes from unannotated emotion text. Most existing methods are pipelined framework, which identifies emotions and extracts causes separately, leading to a drawback of error propagation. Towards this issue, we propose a transition-based model to transform the task into a procedure of parsing-like directed graph construction. The proposed model incrementally generates the directed graph with labeled edges based on a sequence of actions, from which we can recognize emotions with the corresponding causes simultaneously, thereby optimizing separate subtasks jointly and maximizing mutual benefits of tasks interdependently. Experimental results show that our approach achieves the best performance, outperforming the state-of-the-art methods by 6.71% ($p<0.01$) in F1 measure.

Session 7B: Speech and Multimodality-4

CH-SIMS: A Chinese Multimodal Sentiment Analysis Dataset with Fine-grained Annotation of Modality [Website] [PDF]
Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang
9:00–10:00

Previous studies in multimodal sentiment analysis have used limited datasets, which only contain unified multimodal annotations. However, the unified annotations do not always reflect the independent sentiment of single modalities and limit the model to capture the difference between modalities. In this paper, we introduce a Chinese single- and multi-modal sentiment analysis dataset, CH-SIMS, which contains 2,281 refined video segments in the wild with both multimodal and independent unimodal annotations. It allows researchers to study the interaction between modalities or use independent unimodal annotations for unimodal sentiment analysis.Furthermore, we propose a multi-task learning framework based on late fusion as the baseline. Extensive experiments on the CH-SIMS show that our methods achieve state-of-the-art performance and learn more distinctive unimodal representations. The full dataset and codes are available for use at https://github.com/thuiar/MMSA.

Curriculum Pre-training for End-to-End Speech Translation [Website] [PDF]
Chengyi Wang, Yu Wu, Shujie Liu, Ming Zhou, and Zhenglu Yang 9:00–10:00

End-to-end speech translation poses a heavy burden on the encoder because it has to transcribe, understand, and learn cross-lingual semantics simultaneously. To obtain a powerful encoder, traditional methods pre-train it on ASR data to capture speech features. However, we argue that pre-training the encoder only through simple speech recognition is not enough, and high-level linguistic knowledge should be considered. Inspired by this, we propose a curriculum pre-training method that includes an elementary course for transcription learning and two advanced courses for understanding the utterance and mapping words in two languages. The difficulty of these courses is gradually increasing. Experiments show that our curriculum pre-training method leads to significant improvements on En-De and En-Fr speech translation benchmarks.

Improving Disfluency Detection by Self-Training a Self-Attentive Model [Website] [PDF]
Paria Jamshid Lou and Mark Johnson 9:00–10:00

Self-attentive neural syntactic parsers using contextualized word embeddings (e.g. ELMo or BERT) currently produce state-of-the-art results in joint parsing and disfluency detection in speech transcripts. Since the contextualized word embeddings are pre-trained on a large amount of unlabeled data, using additional unlabeled data to train a neural model might seem redundant. However, we show that self-training — a semi-supervised technique for incorporating unlabeled data — sets a new state-of-the-art for the self-attentive parser on disfluency detection, demonstrating that self-training provides benefits orthogonal to the pre-trained contextualized word representations. We also show that ensembling self-trained parsers provides further gains for disfluency detection.

Multimodal Transformer for Multimodal Machine Translation [Website] [PDF]
Shaowei Yao and Xiaojun Wan 9:00–10:00

Multimodal Machine Translation (MMT) aims to introduce information from other modality, generally static images, to improve the translation quality. Previous works propose various incorporation methods, but most of them do not consider the relative importance of multiple modalities. Equally treating all modalities may encode too much useless information from less important modalities. In this paper, we introduce the multimodal self-attention in Transformer to solve the issues above in MMT. The proposed method learns the representation of images based on the text, which avoids encoding irrelevant information in images. Experiments and visualization analysis demonstrate that our model benefits from visual information and substantially outperforms previous works and competitive baselines in terms of various metrics.

Reasoning with Multimodal Sarcastic Tweets via Modeling Cross-Modality Contrast and Semantic Association [Website] [PDF]
Nan Xu, Zhixiong Zeng, and Wenji Mao 9:00–10:00

Sarcasm is a sophisticated linguistic phenomenon to express the opposite of what one really means. With the rapid growth of social media, multimodal sarcastic tweets are widely posted on various social platforms. In multimodal context, sarcasm is no longer a pure linguistic phenomenon, and due to the nature of social media short text, the opposite is more often manifested via cross-modality expressions. Thus traditional text-based methods are insufficient to detect multimodal sarcasm. To reason with multimodal sarcastic tweets, in this paper, we propose a novel method for modeling cross-modality contrast in the associated context. Our method models both cross-modality contrast and semantic association by constructing the Decomposition and Relation Network (namely D&R Net). The decomposition network represents the commonality and discrepancy between image and text, and the relation network models the semantic association in cross-modality context. Experimental results on a public dataset demonstrate the effectiveness of our model in multimodal sarcasm detection.

Sentiment and Emotion help Sarcasm? A Multi-task Learning Framework for Multi-Modal Sarcasm, Sentiment and Emotion Analysis [Website] [PDF]
Dushyant Singh Chauhan, Dhanush S R, Asif Ekbal, and Pushpak Bhattacharyya 9:00–10:00

In this paper, we hypothesize that sarcasm is closely related to sentiment and emotion, and thereby propose a multi-task deep learning framework to solve all these three problems simultaneously in a multi-modal conversational scenario. We, at first, manually annotate the recently released multi-modal MUStARD sarcasm dataset with sentiment and emotion classes, both implicit and explicit. For multi-tasking, we propose two attention mechanisms, viz. Inter-

segment Inter-modal Attention (Ie-Attention) and Intra-segment Inter-modal Attention (Ia-Attention). The main motivation of Ie-Attention is to learn the relationship between the different segments of the sentence across the modalities. In contrast, Ia-Attention focuses within the same segment of the sentence across the modalities. Finally, representations from both the attentions are concatenated and shared across the five classes (i.e., sarcasm, implicit sentiment, explicit sentiment, implicit emotion, explicit emotion) for multi-tasking. Experimental results on the extended version of the MUStARD dataset show the efficacy of our proposed approach for sarcasm detection over the existing state-of-the-art systems. The evaluation also shows that the proposed multi-task framework yields better performance for the primary task, i.e., sarcasm detection, with the help of two secondary tasks, emotion and sentiment analysis.

Towards Emotion-aided Multi-modal Dialogue Act Classification [Website] [PDF]
Tulika Saha, Aditya Patra, Sriparna Saha, and Pushpak Bhattacharyya 9:00–10:00

The task of Dialogue Act Classification (DAC) that purports to capture communicative intent has been studied extensively. But these studies limit themselves to text. Non-verbal features (change of tone, facial expressions etc.) can provide cues to identify DAs, thus stressing the benefit of incorporating multi-modal inputs in the task. Also, the emotional state of the speaker has a substantial effect on the choice of the dialogue act, since conversations are often influenced by emotions. Hence, the effect of emotion too on automatic identification of DAs needs to be studied. In this work, we address the role of *both* multi-modality and emotion recognition (ER) in DAC. DAC and ER help each other by way of multi-task learning. One of the major contributions of this work is a new dataset- multimodal Emotion aware Dialogue Act dataset called EMOTyDA, collected from open-sourced dialogue datasets. To demonstrate the utility of EMOTyDA, we build an attention based (self, inter-modal, inter-task) multi-modal, multi-task Deep Neural Network (DNN) for joint learning of DAs and emotions. We show empirically that multi-modality and multi-tasking achieve better performance of DAC compared to uni-modal and single task DAC variants.

Session 7B: Student Research Workshop

Building a Japanese Typo Dataset from Wikipedia's Revision History [Website][PDF]
Yu Tanaka, Yugo Murawaki, Daisuke Kawahara, and Sadao Kurohashi 9:00–10:00

User generated texts contain many typos for which correction is necessary for NLP systems to work. Although a large number of typo—correction pairs are needed to develop a data-driven typo correction system, no such dataset is available for Japanese. In this paper, we extract over half a million Japanese typo—correction pairs from Wikipedia's revision history. Unlike other languages, Japanese poses unique challenges: (1) Japanese texts are unsegmented so that we cannot simply apply a spelling checker, and (2) the way people inputting kanji logographs results in typos with drastically different surface forms from correct ones. We address them by combining character-based extraction rules, morphological analyzers to guess readings, and various filtering methods. We evaluate the dataset using crowdsourcing and run a baseline seq2seq model for typo correction.

How much complexity does an RNN architecture need to learn syntax-sensitive dependencies? [Website][PDF]
Gantavya Bhatt, Hritik Bansal, Rishubh Singh, and Sumeet Agarwal 9:00–10:00

Long short-term memory (LSTM) networks and their variants are capable of encapsulating long-range dependencies, which is evident from their performance on a variety of linguistic tasks. On the other hand, simple recurrent networks (SRNs), which appear more biologically grounded in terms of synaptic connections, have generally been less successful at capturing long-range dependencies as well as the loci of grammatical errors in an unsupervised setting. In this paper, we seek to develop models that bridge the gap between biological plausibility and linguistic competence. We propose a new architecture, the Decay RNN, which incorporates the decaying nature of neuronal activations and models the excitatory and inhibitory connections in a population of neurons. Besides its biological inspiration, our model also shows competitive performance relative to LSTMs on subject-verb agreement, sentence grammaticality, and language modeling tasks. These results provide some pointers towards probing the nature of the inductive biases required for RNN architectures to model linguistic phenomena successfully.

Unsupervised Multilingual Sentence Embeddings for Parallel Corpus Mining [Website][PDF]
Ivana Kvapilíková, Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Ondřej Bojar 9:00–10:00

Existing models of multilingual sentence embeddings require large parallel data resources which are not available for low-resource languages. We propose a novel unsupervised method to derive multilingual sentence embeddings relying only on monolingual data. We first produce a synthetic parallel corpus using unsupervised machine translation, and use it to fine-tune a pretrained cross-lingual masked language model (XLM) to derive the multilingual sentence representations. The quality of the representations is evaluated on two parallel corpus mining tasks with improvements of up to 22 F1 points over vanilla XLM. In addition, we observe that a single synthetic bilingual corpus is able to improve results for other language pairs.

Demo Session 3A

Time: *12:00–12:45*

EVIDENCEMINER: Textual Evidence Discovery for Life Sciences [Website][PDF]
Xuan Wang, Yingjun Guan, Weili Liu, Aabhas Chauhan, Enyi Jiang, Qi Li, David Liem, Dibakar Sigdel, John Caufield, Peipei Ping, and Jiawei Han

Traditional search engines for life sciences (e.g., PubMed) are designed for document retrieval and do not allow direct retrieval of specific statements. Some of these statements may serve as textual evidence that is key to tasks such as hypothesis generation and new finding validation. We present EVIDENCEMINER, a web-based system that lets users query a natural language statement and automatically retrieves textual evidence from a background corpora for life sciences. EVIDENCEMINER is constructed in a completely automated way without any human effort for training data annotation. It is supported by novel data-driven methods for distantly supervised named entity recognition and open information extraction. The entities and patterns are pre-computed and indexed offline to support fast online evidence retrieval. The annotation results are also highlighted in the original document for better visualization. EVIDENCEMINER also includes analytic functionalities such as the most frequent entity and relation summarization. EVIDENCEMINER can help scientists uncover important research issues, leading to more effective research and more in-depth quantitative analysis. The system of EVIDENCEMINER is available at https://evidenceminer.firebaseapp.com/.

ConvLab-2: An Open-Source Toolkit for Building, Evaluating, and Diagnosing Dialogue Systems [Website][PDF]
Qi Zhu, Zheng Zhang, Yan Fang, Xiang Li, Ryuichi Takanobu, Jinchao Li, Baolin Peng, Jianfeng Gao, Xiaoyan Zhu, and Minlie Huang

We present ConvLab-2, an open-source toolkit that enables researchers to build task-oriented dialogue systems with state-of-the-art models, perform an end-to-end evaluation, and diagnose the weakness of systems. As the successor of ConvLab, ConvLab-2 inherits ConvLab's framework but integrates more powerful dialogue models and supports more datasets. Besides, we have developed an analysis tool and an interactive tool to assist researchers in diagnosing dialogue systems. The analysis tool presents rich statistics and summarizes common mistakes from simulated dialogues, which facilitates error analysis and system improvement. The interactive tool provides an user interface that allows developers to diagnose an assembled dialogue system by interacting with the system and modifying the output of each system component.

Session 8A Overview – Tuesday, July 7, 2020 12:00–13:00

Track A *Computational Social Science and Social Media-6* Abstracts	Analyzing Political Parody in Social Media *Maronikolakis, Sánchez Villegas, Preotiuc-Pietro, and Aletras* [Website] [PDF]	Improving Multimodal Named Entity Recognition via Entity Span Detection with Unified Multimodal Transformer *Yu, Jiang, Yang, and Xia* [Website] [PDF]	Masking Actor Information Leads to Fairer Political Claims Detection *Dayanik and Padó* [Website] [PDF]	Neural Temporal Opinion Modelling for Opinion Prediction on Twitter *Zhu, He, and Zhou* [Website] [PDF]	When do Word Embeddings Accurately Reflect Surveys on our Beliefs About People? *Joseph and Morgan* [Website] [PDF]
	"Who said it, and Why?" Provenance for Natural Language Claims *Zhang, Ives, and Roth* [Website] [PDF]	It Takes Two to Lie: One to Lie, and One to Listen *Peskov, Cheng, Elgohary, Barrow, Danescu-Niculescu-Mizil, and Boyd-Graber* [Website] [PDF]			
Track B *Interpretability and Analysis of Models for NLP-3* Abstracts	Analyzing analytical methods: The case of phonology in neural models of spoken language *Chrupała, Higy, and Alishahi* [Website] [PDF]	Compositionality and Generalization In Emergent Languages *Chaabouni, Kharitonov, Bouchacourt, Dupoux, and Baroni* [Website] [PDF]	ERASER: A Benchmark to Evaluate Rationalized NLP Models *DeYoung, Jain, Rajani, Lehman, Xiong, Socher, and Wallace* [Website] [PDF]	Learning to Faithfully Rationalize by Construction *Jain, Wiegreffe, Pinter, and Wallace* [Website] [PDF]	Make Up Your Mind! Adversarial Generation of Inconsistent Natural Language Explanations *Camburu, Shillingford, Minervini, Lukasiewicz, and Blunsom* [Website] [PDF]
	On the Robustness of Language Encoders against Grammatical Errors *Yin, Long, Meng, and Chang* [Website] [PDF]	Probing for Referential Information in Language Models *Sorodoc, Gulordava, and Boleda* [Website] [PDF]	Roles and Utilization of Attention Heads in Transformer-based Neural Language Models *Jo and Myaeng* [Website] [PDF]	Towards Transparent and Explainable Attention Models *Mohankumar, Nema, Narasimhan, Khapra, Srinivasan, and Ravindran* [Website] [PDF]	Understanding Attention for Text Classification *Sun and Lu* [Website] [PDF]
Track C *Machine Translation-11* Abstracts	Dynamically Adjusting Transformer Batch Size by Monitoring Gradient Direction Change *Xu, Genabith, Xiong, and Liu* [Website] [PDF]	Geometry-aware domain adaptation for unsupervised alignment of word embeddings *Jawanpuria, Meghwanshi, and Mishra* [Website] [PDF]	Modeling Word Formation in English—German Neural Machine Translation *Weller-Di Marco and Fraser* [Website] [PDF]	On Exposure Bias, Hallucination and Domain Shift in Neural Machine Translation *Wang and Sennrich* [Website] [PDF]	Successfully Applying the Stabilized Lottery Ticket Hypothesis to the Transformer Architecture *Brix, Bahar, and Ney* [Website] [PDF]
Track D *Question Answering-5* Abstracts	A Self-Training Method for Machine Reading Comprehension with Soft Evidence Extraction *Niu, Jiao, Zhou, Yao, and Huang* [Website] [PDF]	[TACL] Break It Down: A Question Understanding Benchmark *Wolfson, Geva, Gupta, Goldberg, Gardner, Deutch, and Berant* [Website] [PDF]	Clinical Reading Comprehension: A Thorough Analysis of the emrQA Dataset *Yue, Jimenez Gutierrez, and Sun* [Website] [PDF]	DeFormer: Decomposing Pre-trained Transformers for Faster Question Answering *Cao, Trivedi, Balasubramanian, and Balasubramanian* [Website] [PDF]	Graph-to-Tree Learning for Solving Math Word Problems *Zhang, Wang, Lee, Bin, Wang, Shao, and Lim* [Website] [PDF]

	Improving Multi-hop Question Answering over Knowledge Graphs using Knowledge Base Embeddings *Saxena, Tripathi, and Talukdar* [Website][PDF]	Template-Based Question Generation from Retrieved Sentences for Improved Unsupervised Question Answering *Fabbri, Ng, Wang, Nallapati, and Xiang* [Website][PDF]	Unsupervised Alignment-based Iterative Evidence Retrieval for Multi-hop Question Answering *Yadav, Bethard, and Surdeanu* [Website][PDF]		
Track E *Resources and Evaluation-7* Abstracts	A Corpus for Large-Scale Phonetic Typology *Salesky, Chodroff, Pimentel, Wiesner, Cotterell, Black, and Eisner* [Website][PDF]	Dscorer: A Fast Evaluation Metric for Discourse Representation Structure Parsing *Liu, Cohen, and Lapata* [Website][PDF]	MATINF: A Jointly Labeled Large-Scale Dataset for Classification, Question Answering and Summarization *Xu, Pei, Wu, Liu, and Li* [Website][PDF]	MIND: A Large-scale Dataset for News Recommendation *Wu, Qiao, Chen, Wu, Qi, Lian, Liu, Xie, Gao, Wu, and Zhou* [Website][PDF]	ParaCrawl: Web-Scale Acquisition of Parallel Corpora *Bañón, Chen, Haddow, Heafield, Hoang, Esplà-Gomis, Forcada, Kamran, Kirefu, Koehn, Ortiz Rojas, Pla Sempere, Ramírez-Sánchez, Sarrías, Strelec, Thompson, Waites, Wiggins, and Zaragoza* [Website][PDF]
Track F *Lexical-6* Abstracts	Adaptive Compression of Word Embeddings *Kim, Kim, and Lee* [Website][PDF]	Analysing Lexical Semantic Change with Contextualised Word Representations *Giulianelli, Del Tredici, and Fernández* [Website][PDF]	Autoencoding Keyword Correlation Graph for Document Clustering *Chiu, Sahu, Thomas, Sengupta, and Mahdy* [Website][PDF]	Autoencoding Pixies: Amortised Variational Inference with Graph Convolutions for Functional Distributional Semantics *Emerson* [Website][PDF]	BERTRAM: Improved Word Embeddings Have Big Impact on Contextualized Model Performance *Schick and Schütze* [Website][PDF]
	CluBERT: A Cluster-Based Approach for Learning Sense Distributions in Multiple Languages *Pasini, Scozzafava, and Scarlini* [Website][PDF]	Hypernymy Detection for Low-Resource Languages via Meta Learning *Yu, Han, Zhang, and Ng* [Website][PDF]	Investigating Word-Class Distributions in Word Vector Spaces *Sasano and Korhonen* [Website][PDF]		
Track G *Sentence Level-4* Abstracts	AMR Parsing with Latent Structural Information *Zhou, Zhang, Ji, and Tang* [Website][PDF]	TaPas: Weakly Supervised Table Parsing via Pre-training *Herzig, Nowak, Müller, Piccinno, and Eisenschlos* [Website][PDF]			
Track H *Sentiment Analysis, Stylistic Analysis, and Argument Mining-5* Abstracts	Adversarial and Domain-Aware BERT for Cross-Domain Sentiment Analysis *Du, Sun, Wang, Qi, and Liao* [Website][PDF]	Don't Eclipse Your Arts Due to Small Discrepancies: Boundary Repositioning with a Pointer Network for Aspect Extraction *Wei, Hong, Zou, Cheng, and YAO* [Website][PDF]	Relational Graph Attention Network for Aspect-based Sentiment Analysis *Wang, Shen, Yang, Quan, and Wang* [Website][PDF]	SentiBERT: A Transferable Transformer-Based Architecture for Compositional Sentiment Semantics *Yin, Meng, and Chang* [Website][PDF]	Target Inference in Argument Conclusion Generation *Alshomary, Syed, Potthast, and Wachsmuth* [Website][PDF]

Track I *Student Research Workshop* Abstracts	How much complexity does an RNN architecture need to learn syntax-sensitive dependencies? *Bhatt, Bansal, Singh, and Agarwal* [Website][PDF]	Logical Inferences with Comparatives and Generalized Quantifiers *Haruta, Mineshima, and Bekki* [Website][PDF]	Unsupervised Multilingual Sentence Embeddings for Parallel Corpus Mining *Kvapilíková, Artetxe, Labaka, Agirre, and Bojar* [Website][PDF]	Enhancing Word Embeddings with Knowledge Extracted from Lexical Resources *Biesialska, Rafieian, and Costa-jussà* [Website][PDF]	
Track J *Tagging, Chunking and Parsing-3* Abstracts	A Span-based Linearization for Constituent Trees *Wei, Wu, and Lan* [Website][PDF]	[CL] Abstract Syntax as Interlingua: Scaling Up the Grammatical Framework from Controlled Languages to Robust Pipelines *Ranta, Angelov, Gruzitis, and Kolachina* [Website][PDF]	Do Neural Language Models Show Preferences for Syntactic Formalisms? *Kulmizev, Ravishankar, Abdou, and Nivre* [Website][PDF]	Enriched In-Order Linearization for Faster Sequence-to-Sequence Constituent Parsing *Fernández-González and Gómez-Rodríguez* [Website][PDF]	Exact yet Efficient Graph Parsing, Bi-directional Locality and the Constructivist Hypothesis *Ye and Sun* [Website][PDF]
	Max-Margin Incremental CCG Parsing *Stanojević and Steedman* [Website][PDF]	Neural Reranking for Dependency Parsing: An Evaluation *Do and Rehbein* [Website][PDF]			

Session 8A Details

Session 8A: Computational Social Science and Social Media-6

Analyzing Political Parody in Social Media [Website][PDF]
Antonios Maronikolakis, Danae Sánchez Villegas, Daniel Preotiuc-Pietro, and Nikolaos Aletras 12:00–13:00

Parody is a figurative device used to imitate an entity for comedic or critical purposes and represents a widespread phenomenon in social media through many popular parody accounts. In this paper, we present the first computational study of parody. We introduce a new publicly available data set of tweets from real politicians and their corresponding parody accounts. We run a battery of supervised machine learning models for automatically detecting parody tweets with an emphasis on robustness by testing on tweets from accounts unseen in training, across different genders and across countries. Our results show that political parody tweets can be predicted with an accuracy up to 90%. Finally, we identify the markers of parody through a linguistic analysis. Beyond research in linguistics and political communication, accurately and automatically detecting parody is important to improving fact checking for journalists and analytics such as sentiment analysis through filtering out parodical utterances.

Improving Multimodal Named Entity Recognition via Entity Span Detection with Unified Multimodal Transformer [Website][PDF]
Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia 12:00–13:00

In this paper, we study Multimodal Named Entity Recognition (MNER) for social media posts. Existing approaches for MNER mainly suffer from two drawbacks: (1) despite generating word-aware visual representations, their word representations are insensitive to the visual context; (2) most of them ignore the bias brought by the visual context. To tackle the first issue, we propose a multimodal interaction module to obtain both image-aware word representations and word-aware visual representations. To alleviate the visual bias, we further propose to leverage purely text-based entity span detection as an auxiliary module, and design a Unified Multimodal Transformer to guide the final predictions with the entity span predictions. Experiments show that our unified approach achieves the new state-of-the-art performance on two benchmark datasets.

Masking Actor Information Leads to Fairer Political Claims Detection [Website][PDF]
Erenay Dayanik and Sebastian Padó 12:00–13:00

A central concern in Computational Social Sciences (CSS) is fairness: where the role of NLP is to scale up text analysis to large corpora, the quality of automatic analyses should be as independent as possible of textual properties. We analyze the performance of a state-of-the-art neural model on the task of political claims detection (i.e., the identification of forward-looking statements made by political actors) and identify a strong frequency bias: claims made by frequent actors are recognized better. We propose two simple debiasing methods which mask proper names and pronouns during training of the model, thus removing personal information bias. We find that (a) these methods significantly decrease frequency bias while keeping the overall performance stable; and (b) the resulting models improve when evaluated in an out-of-domain setting.

Neural Temporal Opinion Modelling for Opinion Prediction on Twitter [Website][PDF]
Lixing Zhu, Yulan He, and Deyu Zhou 12:00–13:00

Opinion prediction on Twitter is challenging due to the transient nature of tweet content and neighbourhood context. In this paper, we model users' tweet posting behaviour as a temporal point process to jointly predict the posting time and the stance label of the next tweet given a user's historical tweet sequence and tweets posted by their neighbours. We design a topic-driven attention mechanism to capture the dynamic topic shifts in the neighbourhood context. Experimental results show that the proposed model predicts both the posting time and the stance labels of future tweets more accurately compared to a number of competitive baselines.

When do Word Embeddings Accurately Reflect Surveys on our Beliefs About People? [Website][PDF]
Kenneth Joseph and Jonathan Morgan 12:00–13:00

Social biases are encoded in word embeddings. This presents a unique opportunity to study society historically and at scale, and a unique danger when embeddings are used in downstream applications. Here, we investigate the extent to which publicly-available word embeddings accurately reflect beliefs about certain kinds of people as measured via traditional survey methods. We find that biases found in word embeddings do, on average, closely mirror survey data across seventeen dimensions of social meaning. However, we also find that biases in embeddings are much more reflective of survey data for some dimensions of meaning (e.g. gender) than others (e.g. race), and that we can be highly confident that embedding-based measures reflect survey data only for the most salient biases.

"Who said it, and Why?" Provenance for Natural Language Claims [Website][PDF]
Yi Zhang, Zachary Ives, and Dan Roth 12:00–13:00

In an era where generating content and publishing it is so easy, we are bombarded with information and are exposed to all kinds of claims, some of which do not always rank high on the truth scale. This paper suggests that the key to a longer-term, holistic, and systematic approach to navigating this information pollution is capturing the provenance of claims. To do that, we develop a formal definition of provenance graph for a given natural language claim, aiming to understand where the claim may come from and how it has evolved. To construct the graph, we model provenance

inference, formulated mainly as an information extraction task and addressed via a textual entailment model. We evaluate our approach using two benchmark datasets, showing initial success in capturing the notion of provenance and its effectiveness on the application of claim verification.

It Takes Two to Lie: One to Lie, and One to Listen [Website] [PDF]

Denis Peskov, Benny Cheng, Ahmed Elgohary, Joe Barrow, Cristian Danescu-Niculescu-Mizil, and Jordan Boyd-Graber 12:00–13:00

Trust is implicit in many online text conversations—striking up new friendships, or asking for tech support. But trust can be betrayed through deception. We study the language and dynamics of deception in the negotiation-based game Diplomacy, where seven players compete for world domination by forging and breaking alliances with each other. Our study with players from the Diplomacy community gathers 17,289 messages annotated by the sender for their intended truthfulness and by the receiver for their perceived truthfulness. Unlike existing datasets, this captures deception in long-lasting relationships, where the interlocutors strategically combine truth with lies to advance objectives. A model that uses power dynamics and conversational contexts can predict when a lie occurs nearly as well as human players.

Session 8A: Interpretability and Analysis of Models for NLP-3

Analyzing analytical methods: The case of phonology in neural models of spoken language [Website][PDF]
Grzegorz Chrupała, Bertrand Higy, and Afra Alishahi 12:00–13:00

Given the fast development of analysis techniques for NLP and speech processing systems, few systematic studies have been conducted to compare the strengths and weaknesses of each method. As a step in this direction we study the case of representations of phonology in neural network models of spoken language. We use two commonly applied analytical techniques, diagnostic classifiers and representational similarity analysis, to quantify to what extent neural activation patterns encode phonemes and phoneme sequences. We manipulate two factors that can affect the outcome of analysis. First, we investigate the role of learning by comparing neural activations extracted from trained versus randomly-initialized models. Second, we examine the temporal scope of the activations by probing both local activations corresponding to a few milliseconds of the speech signal, and global activations pooled over the whole utterance. We conclude that reporting analysis results with randomly initialized models is crucial, and that global-scope methods tend to yield more consistent and interpretable results and we recommend their use as a complement to local-scope diagnostic methods.

Compositionality and Generalization In Emergent Languages [Website][PDF]
Rahma Chaabouni, Eugene Kharitonov, Diane Bouchacourt, Emmanuel Dupoux, and Marco Baroni
12:00–13:00

Natural language allows us to refer to novel composite concepts by combining expressions denoting their parts according to systematic rules, a property known as compositionality. In this paper, we study whether the language emerging in deep multi-agent simulations possesses a similar ability to refer to novel primitive combinations, and whether it accomplishes this feat by strategies akin to human-language compositionality. Equipped with new ways to measure compositionality in emergent languages inspired by disentanglement in representation learning, we establish three main results: First, given sufficiently large input spaces, the emergent language will naturally develop the ability to refer to novel composite concepts. Second, there is no correlation between the degree of compositionality of an emergent language and its ability to generalize. Third, while compositionality is not necessary for generalization, it provides an advantage in terms of language transmission: The more compositional a language is, the more easily it will be picked up by new learners, even when the latter differ in architecture from the original agents. We conclude that compositionality does not arise from simple generalization pressure, but if an emergent language does chance upon it, it will be more likely to survive and thrive.

ERASER: A Benchmark to Evaluate Rationalized NLP Models [Website][PDF]
Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace 12:00–13:00

State-of-the-art models in NLP are now predominantly based on deep neural networks that are opaque in terms of how they come to make predictions. This limitation has increased interest in designing more interpretable deep models for NLP that reveal the 'reasoning' behind model outputs. But work in this direction has been conducted on different datasets and tasks with correspondingly unique aims and metrics; this makes it difficult to track progress. We propose the **E**valuating **R**ationales **A**nd **S**imple **E**nglish **R**easoning (**ERASER** a benchmark to advance research on interpretable models in NLP. This benchmark comprises multiple datasets and tasks for which human annotations of "rationales" (supporting evidence) have been collected. We propose several metrics that aim to capture how well the rationales provided by models align with human rationales, and also how *faithful* these rationales are (i.e., the degree to which provided rationales influenced the corresponding predictions). Our hope is that releasing this benchmark facilitates progress on designing more interpretable NLP systems. The benchmark, code, and documentation are available at `https://www.eraserbenchmark.com/`

Learning to Faithfully Rationalize by Construction [Website][PDF]
Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C. Wallace 12:00–13:00

In many settings it is important for one to be able to understand why a model made a particular prediction. In NLP this often entails extracting snippets of an input text 'responsible for' corresponding model output; when such a snippet comprises tokens that indeed informed the model's prediction, it is a faithful explanation. In some settings, faithfulness may be critical to ensure transparency. Lei et al. (2016) proposed a model to produce faithful rationales for neural text classification by defining independent snippet extraction and prediction modules. However, the discrete selection over input tokens performed by this method complicates training, leading to high variance and requiring careful hyperparameter tuning. We propose a simpler variant of this approach that provides faithful explanations by construction. In our scheme, named FRESH, arbitrary feature importance scores (e.g., gradients from a trained model) are used to induce binary labels over token inputs, which an extractor can be trained to predict. An independent classifier module is then trained exclusively on snippets provided by the extractor; these snippets thus constitute faithful explanations, even if the classifier is arbitrarily complex. In both automatic and manual evaluations we find that variants of this simple framework yield predictive performance superior to 'end-to-end' approaches, while being more general and easier to train. Code is available at https://github.com/successar/FRESH.

Make Up Your Mind! Adversarial Generation of Inconsistent Natural Language Explanations [Website][PDF]
Oana-Maria Camburu, Brendan Shillingford, Pasquale Minervini, Thomas Lukasiewicz, and Phil Blunsom 12:00–13:00

To increase trust in artificial intelligence systems, a promising research direction consists of designing neural models capable of generating natural language explanations for their predictions. In this work, we show that such models are nonetheless prone to generating mutually inconsistent explanations, such as "Because there is a dog in the image." and "Because there is no dog in the [same] image.", exposing flaws in either the decision-making process of the model or in the generation of the explanations. We introduce a simple yet effective adversarial framework for sanity checking models against the generation of inconsistent natural language explanations. Moreover, as part of the framework, we address the problem of adversarial attacks with full target sequences, a scenario that was not previously addressed in sequence-to-sequence attacks. Finally, we apply our framework on a state-of-the-art neural natural language inference model that provides natural language explanations for its predictions. Our framework shows that this model is capable of generating a significant number of inconsistent explanations.

On the Robustness of Language Encoders against Grammatical Errors [Website][PDF]
Fan Yin, Quanyu Long, Tao Meng, and Kai-Wei Chang 12:00–13:00

We conduct a thorough study to diagnose the behaviors of pre-trained language encoders (ELMo, BERT, and RoBERTa) when confronted with natural grammatical errors. Specifically, we collect real grammatical errors from non-native speakers and conduct adversarial attacks to simulate these errors on clean text data. We use this approach to facilitate debugging models on downstream applications. Results confirm that the performance of all tested models is affected but the degree of impact varies. To interpret model behaviors, we further design a linguistic acceptability task to reveal their abilities in identifying ungrammatical sentences and the position of errors. We find that fixed contextual encoders with a simple classifier trained on the prediction of sentence correctness are able to locate error positions. We also design a cloze test for BERT and discover that BERT captures the interaction between errors and specific tokens in context. Our results shed light on understanding the robustness and behaviors of language encoders against grammatical errors.

Probing for Referential Information in Language Models [Website][PDF]
Ionut-Teodor Sorodoc, Kristina Gulordava, and Gemma Boleda 12:00–13:00

Language models keep track of complex information about the preceding context – including, e.g., syntactic relations in a sentence. We investigate whether they also capture information beneficial for resolving pronominal anaphora in English. We analyze two state of the art models with LSTM and Transformer architectures, via probe tasks and analysis on a coreference annotated corpus. The Transformer outperforms the LSTM in all analyses. Our results suggest that language models are more successful at learning grammatical constraints than they are at learning truly referential information, in the sense of capturing the fact that we use language to refer to entities in the world. However, we find traces of the latter aspect, too.

Roles and Utilization of Attention Heads in Transformer-based Neural Language Models [Website][PDF]
Jae-young Jo and Sung-Hyon Myaeng 12:00–13:00

Sentence encoders based on the transformer architecture have shown promising results on various natural language tasks. The main impetus lies in the pre-trained neural language models that capture long-range dependencies among words, owing to multi-head attention that is unique in the architecture. However, little is known for how linguistic properties are processed, represented, and utilized for downstream tasks among hundreds of attention heads inside the pre-trained transformer-based model. For the initial goal of examining the roles of attention heads in handling a set of linguistic features, we conducted a set of experiments with ten probing tasks and three downstream tasks on four pre-trained transformer families (GPT, GPT2, BERT, and ELECTRA). Meaningful insights are shown through the lens of heat map visualization and utilized to propose a relatively simple sentence representation method that takes advantage of most influential attention heads, resulting in additional performance improvements on the downstream tasks.

Towards Transparent and Explainable Attention Models [Website][PDF]
Akash Kumar Mohankumar, Preksha Nema, Sharan Narasimhan, Mitesh M. Khapra, Balaji Vasan Srinivasan, and Balaraman Ravindran 12:00–13:00

Recent studies on interpretability of attention distributions have led to notions of faithful and plausible explanations for a model's predictions. Attention distributions can be considered a faithful explanation if a higher attention weight implies a greater impact on the model's prediction. They can be considered a plausible explanation if they provide a human-understandable justification for the model's predictions. In this work, we first explain why current attention mechanisms in LSTM based encoders can neither provide a faithful nor a plausible explanation of the model's predictions. We observe that in LSTM based encoders the hidden representations at different time-steps are very similar to each other (high conicity) and attention weights in these situations do not carry much meaning because even a random permutation of the attention weights does not affect the model's predictions. Based on experiments on a wide variety of tasks and datasets, we observe attention distributions often attribute the model's predictions to unimportant words such as punctuation and fail to offer a plausible explanation for the predictions. To make attention mechanisms more faithful and plausible, we propose a modified LSTM cell with a diversity-driven training objective that ensures that the hidden representations learned at different time steps are diverse. We show that the resulting attention distributions offer more transparency as they (i) provide a more precise importance ranking of the hidden states (ii) are better indicative of words important for the model's predictions (iii) correlate better with gradient-based attribution methods. Human evaluations indicate that the attention distributions learned by our model offer a plausible explanation of the model's predictions. Our code has been made publicly available at https://github.com/akashkm99/Interpretable-Attention

Understanding Attention for Text Classification [Website][PDF]
Xiaobing Sun and Wei Lu 12:00–13:00

Attention has been proven successful in many natural language processing (NLP) tasks. Recently, many researchers started to investigate the interpretability of attention on NLP tasks. Many existing approaches focused on examining whether the local attention weights could reflect the importance of input representations. In this work, we present a study on understanding the internal mechanism of attention by looking into the gradient update process, checking its behavior when approaching a local minimum during training. We propose to analyze for each word token the following two quantities: its polarity score and its attention score, where the latter is a global assessment on the token's significance. We discuss conditions under which the attention mechanism may become more (or less) interpretable, and show how the interplay between the two quantities can contribute towards model performance.

Session 8A: Machine Translation-11

Dynamically Adjusting Transformer Batch Size by Monitoring Gradient Direction Change [Website][PDF]
Hongfei Xu, Josef van Genabith, Deyi Xiong, and Qiuhui Liu 12:00–13:00

The choice of hyper-parameters affects the performance of neural models. While much previous research (Sutskever et al., 2013; Duchi et al., 2011; Kingma and Ba, 2015) focuses on accelerating convergence and reducing the effects of the learning rate, comparatively few papers concentrate on the effect of batch size. In this paper, we analyze how increasing batch size affects gradient direction, and propose to evaluate the stability of gradients with their angle change. Based on our observations, the angle change of gradient direction first tends to stabilize (i.e. gradually decrease) while accumulating mini-batches, and then starts to fluctuate. We propose to automatically and dynamically determine batch sizes by accumulating gradients of mini-batches and performing an optimization step at just the time when the direction of gradients starts to fluctuate. To improve the efficiency of our approach for large models, we propose a sampling approach to select gradients of parameters sensitive to the batch size. Our approach dynamically determines proper and efficient batch sizes during training. In our experiments on the WMT 14 English to German and English to French tasks, our approach improves the Transformer with a fixed 25k batch size by +0.73 and +0.82 BLEU respectively.

Geometry-aware domain adaptation for unsupervised alignment of word embeddings [Website][PDF]
Pratik Jawanpuria, Mayank Meghwanshi, and Bamdev Mishra 12:00–13:00

We propose a novel manifold based geometric approach for learning unsupervised alignment of word embeddings between the source and the target languages. Our approach formulates the alignment learning problem as a domain adaptation problem over the manifold of doubly stochastic matrices. This viewpoint arises from the aim to align the second order information of the two language spaces. The rich geometry of the doubly stochastic manifold allows to employ efficient Riemannian conjugate gradient algorithm for the proposed formulation. Empirically, the proposed approach outperforms state-of-the-art optimal transport based approach on the bilingual lexicon induction task across several language pairs. The performance improvement is more significant for distant language pairs.

Modeling Word Formation in English—German Neural Machine Translation [Website][PDF]
Marion Weller-Di Marco and Alexander Fraser 12:00–13:00

This paper studies strategies to model word formation in NMT using rich linguistic information, namely a word segmentation approach that goes beyond splitting into substrings by considering fusional morphology. Our linguistically sound segmentation is combined with a method for target-side inflection to accommodate modeling word formation. The best system variants employ source-side morphological analysis and model complex target-side words, improving over a standard system.

On Exposure Bias, Hallucination and Domain Shift in Neural Machine Translation [Website][PDF]
Chaojun Wang and Rico Sennrich 12:00–13:00

The standard training algorithm in neural machine translation (NMT) suffers from exposure bias, and alternative algorithms have been proposed to mitigate this. However, the practical impact of exposure bias is under debate. In this paper, we link exposure bias to another well-known problem in NMT, namely the tendency to generate hallucinations under domain shift. In experiments on three datasets with multiple test domains, we show that exposure bias is partially to blame for hallucinations, and that training with Minimum Risk Training, which avoids exposure bias, can mitigate this. Our analysis explains why exposure bias is more problematic under domain shift, and also links exposure bias to the beam search problem, i.e. performance deterioration with increasing beam size. Our results provide a new justification for methods that reduce exposure bias: even if they do not increase performance on in-domain test sets, they can increase model robustness to domain shift.

Successfully Applying the Stabilized Lottery Ticket Hypothesis to the Transformer Architecture [Website][PDF]
Christopher Brix, Parnia Bahar, and Hermann Ney 12:00–13:00

Sparse models require less memory for storage and enable a faster inference by reducing the necessary number of FLOPs. This is relevant both for time-critical and on-device computations using neural networks. The stabilized lottery ticket hypothesis states that networks can be pruned after none or few training iterations, using a mask computed based on the unpruned converged model. On the transformer architecture and the WMT 2014 English-to-German and English-to-French tasks, we show that stabilized lottery ticket pruning performs similar to magnitude pruning for sparsity levels of up to 85%, and propose a new combination of pruning techniques that outperforms all other techniques for even higher levels of sparsity. Furthermore, we confirm that the parameter's initial sign and not its specific value is the primary factor for successful training, and show that magnitude pruning cannot be used to find winning lottery tickets.

Session 8A: Question Answering-5

A Self-Training Method for Machine Reading Comprehension with Soft Evidence Extraction [Website][PDF]
Yilin Niu, Fangkai Jiao, Mantong Zhou, Ting Yao, jingfang xu jingfang, and Minlie Huang 12:00–13:00

Neural models have achieved great success on machine reading comprehension (MRC), many of which typically consist of two components: an evidence extractor and an answer predictor. The former seeks the most relevant information from a reference text, while the latter is to locate or generate answers from the extracted evidence. Despite the importance of evidence labels for training the evidence extractor, they are not cheaply accessible, particularly in many non-extractive MRC tasks such as YES/NO question answering and multi-choice MRC. To address this problem, we present a Self-Training method (STM), which supervises the evidence extractor with auto-generated evidence labels in an iterative process. At each iteration, a base MRC model is trained with golden answers and noisy evidence labels. The trained model will predict pseudo evidence labels as extra supervision in the next iteration. We evaluate STM on seven datasets over three MRC tasks. Experimental results demonstrate the improvement on existing MRC models, and we also analyze how and why such a self-training method works in MRC.

[TACL] Break It Down: A Question Understanding Benchmark [Website][PDF]
Tomer Wolfson, Mor Geva, Ankit Gupta, Yoav Goldberg, Matt Gardner, Daniel Deutch, and Jonathan Berant 12:00–13:00

Understanding natural language questions entails the ability to break down a question into the requisite steps for computing its answer. In this work, we introduce a Question Decomposition Meaning Representation (QDMR) for questions. QDMR constitutes the ordered list of steps, expressed through natural language, that are necessary for answering a question. We develop a crowdsourcing pipeline, showing that quality QDMRs can be annotated at scale, and release the Break dataset, containing over 83K pairs of questions and their QDMRs.

Clinical Reading Comprehension: A Thorough Analysis of the emrQA Dataset [Website][PDF]
Xiang Yue, Bernal Jimenez Gutierrez, and Huan Sun 12:00–13:00

Machine reading comprehension has made great progress in recent years owing to large-scale annotated datasets. In the clinical domain, however, creating such datasets is quite difficult due to the domain expertise required for annotation. Recently, Pampari et al. (EMNLP'18) tackled this issue by using expert-annotated question templates and existing i2b2 annotations to create emrQA, the first large-scale dataset for question answering (QA) based on clinical notes. In this paper, we provide an in-depth analysis of this dataset and the clinical reading comprehension (CliniRC) task. From our qualitative analysis, we find that (i) emrQA answers are often incomplete, and (ii) emrQA questions are often answerable without using domain knowledge. From our quantitative experiments, surprising results include that (iii) using a small sampled subset (5%-20%), we can obtain roughly equal performance compared to the model trained on the entire dataset, (iv) this performance is close to human expert's performance, and (v) BERT models do not beat the best performing base model. Following our analysis of the emrQA, we further explore two desired aspects of CliniRC systems: the ability to utilize clinical domain knowledge and to generalize to unseen questions and contexts. We argue that both should be considered when creating future datasets.

DeFormer: Decomposing Pre-trained Transformers for Faster Question Answering [Website][PDF]
Qingqing Cao, Harsh Trivedi, Aruna Balasubramanian, and Niranjan Balasubramanian 12:00–13:00

Transformer-based QA models use input-wide self-attention – i.e. across both the question and the input passage – at all layers, causing them to be slow and memory-intensive. It turns out that we can get by without input-wide self-attention at all layers, especially in the lower layers. We introduce DeFormer, a decomposed transformer, which substitutes the full self-attention with question-wide and passage-wide self-attentions in the lower layers. This allows for question-independent processing of the input text representations, which in turn enables pre-computing passage representations reducing runtime compute drastically. Furthermore, because DeFormer is largely similar to the original model, we can initialize DeFormer with the pre-training weights of a standard transformer, and directly fine-tune on the target QA dataset. We show DeFormer versions of BERT and XLNet can be used to speed up QA by over 4.3x and with simple distillation-based losses they incur only a 1% drop in accuracy. We open source the code at https://github.com/StonyBrookNLP/deformer.

Graph-to-Tree Learning for Solving Math Word Problems [Website][PDF]
Jipeng Zhang, Lei Wang, Roy Ka-Wei Lee, Yi Bin, Yan Wang, Jie Shao, and Ee-Peng Lim 12:00–13:00

While the recent tree-based neural models have demonstrated promising results in generating solution expression for the math word problem (MWP), most of these models do not capture the relationships and order information among the quantities well. This results in poor quantity representations and incorrect solution expressions. In this paper, we propose Graph2Tree, a novel deep learning architecture that combines the merits of the graph-based encoder and tree-based decoder to generate better solution expressions. Included in our Graph2Tree framework are two graphs, namely the Quantity Cell Graph and Quantity Comparison Graph, which are designed to address limitations of existing methods by effectively representing the relationships and order information among the quantities in MWPs. We conduct extensive experiments on two available datasets. Our experiment results show that Graph2Tree outperforms the state-of-the-art baselines on two benchmark datasets significantly. We also discuss case studies and empirically examine Graph2Tree's effectiveness in translating the MWP text into solution expressions.

Improving Multi-hop Question Answering over Knowledge Graphs using Knowledge Base Embeddings [Website][PDF]
Apoorv Saxena, Aditay Tripathi, and Partha Talukdar 12:00–13:00

Knowledge Graphs (KG) are multi-relational graphs consisting of entities as nodes and relations among them as typed edges. Goal of the Question Answering over KG (KGQA) task is to answer natural language queries posed over the KG. Multi-hop KGQA requires reasoning over multiple edges of the KG to arrive at the right answer. KGs are often incomplete with many missing links, posing additional challenges for KGQA, especially for multi-hop KGQA. Recent research on multi-hop KGQA has attempted to handle KG sparsity using relevant external text, which isn't always readily available. In a separate line of research, KG embedding methods have been proposed to reduce KG sparsity by performing missing link prediction. Such KG embedding methods, even though highly relevant, have not been explored for multi-hop KGQA so far. We fill this gap in this paper and propose EmbedKGQA. EmbedKGQA is particularly effective in performing multi-hop KGQA over sparse KGs. EmbedKGQA also relaxes the requirement of answer selection from a pre-specified neighborhood, a sub-optimal constraint enforced by previous multi-hop KGQA methods. Through extensive experiments on multiple benchmark datasets, we demonstrate EmbedKGQA's effectiveness over other state-of-the-art baselines.

Template-Based Question Generation from Retrieved Sentences for Improved Unsupervised Question Answering [Website][PDF]

Alexander Fabbri, Patrick Ng, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang 12:00–13:00

Question Answering (QA) is in increasing demand as the amount of information available online and the desire for quick access to this content grows. A common approach to QA has been to fine-tune a pretrained language model on a task-specific labeled dataset. This paradigm, however, relies on scarce, and costly to obtain, large-scale human-labeled data. We propose an unsupervised approach to training QA models with generated pseudo-training data. We show that generating questions for QA training by applying a simple template on a related, retrieved sentence rather than the original context sentence improves downstream QA performance by allowing the model to learn more complex context-question relationships. Training a QA model on this data gives a relative improvement over a previous unsupervised model in F1 score on the SQuAD dataset by about 14%, and 20% when the answer is a named entity, achieving state-of-the-art performance on SQuAD for unsupervised QA.

Unsupervised Alignment-based Iterative Evidence Retrieval for Multi-hop Question Answering [Website][PDF]

Vikas Yadav, Steven Bethard, and Mihai Surdeanu 12:00–13:00

Evidence retrieval is a critical stage of question answering (QA), necessary not only to improve performance, but also to explain the decisions of the QA method. We introduce a simple, fast, and unsupervised iterative evidence retrieval method, which relies on three ideas: (a) an unsupervised alignment approach to soft-align questions and answers with justification sentences using only GloVe embeddings, (b) an iterative process that reformulates queries focusing on terms that are not covered by existing justifications, which (c) stops when the terms in the given question and candidate answers are covered by the retrieved justifications. Despite its simplicity, our approach outperforms all the previous methods (including supervised methods) on the evidence selection task on two datasets: MultiRC and QASC. When these evidence sentences are fed into a RoBERTa answer classification component, we achieve state-of-the-art QA performance on these two datasets.

Session 8A: Resources and Evaluation-7

A Corpus for Large-Scale Phonetic Typology [Website][PDF]
Elizabeth Salesky, Eleanor Chodroff, Tiago Pimentel, Matthew Wiesner, Ryan Cotterell, Alan W Black, and Jason Eisner 12:00–13:00

A major hurdle in data-driven research on typology is having sufficient data in many languages to draw meaningful conclusions. We present VoxClamantis v1.0, the first large-scale corpus for phonetic typology, with aligned segments and estimated phoneme-level labels in 690 readings spanning 635 languages, along with acoustic-phonetic measures of vowels and sibilants. Access to such data can greatly facilitate investigation of phonetic typology at a large scale and across many languages. However, it is non-trivial and computationally intensive to obtain such alignments for hundreds of languages, many of which have few to no resources presently available. We describe the methodology to create our corpus, discuss caveats with current methods and their impact on the utility of this data, and illustrate possible research directions through a series of case studies on the 48 highest-quality readings. Our corpus and scripts are publicly available for non-commercial use at https://voxclamantisproject.github.io.

Dscorer: A Fast Evaluation Metric for Discourse Representation Structure Parsing [Website][PDF]
Jiangming Liu, Shay B. Cohen, and Mirella Lapata 12:00–13:00

Discourse representation structures (DRSs) are scoped semantic representations for texts of arbitrary length. Evaluating the accuracy of predicted DRSs plays a key role in developing semantic parsers and improving their performance. DRSs are typically visualized as boxes which are not straightforward to process automatically. Counter transforms DRSs to clauses and measures clause overlap by searching for variable mappings between two DRSs. However, this metric is computationally costly (with respect to memory and CPU time) and does not scale with longer texts. We introduce Dscorer, an efficient new metric which converts box-style DRSs to graphs and then measures the overlap of n-grams. Experiments show that Dscorer computes accuracy scores that are correlated with Counter at a fraction of the time.

MATINF: A Jointly Labeled Large-Scale Dataset for Classification, Question Answering and Summarization [Website][PDF]
Canwen Xu, Jiaxin Pei, Hongtao Wu, Yiyu Liu, and Chenliang Li 12:00–13:00

Recently, large-scale datasets have vastly facilitated the development in nearly all domains of Natural Language Processing. However, there is currently no cross-task dataset in NLP, which hinders the development of multi-task learning. We propose MATINF, the first jointly labeled large-scale dataset for classification, question answering and summarization. MATINF contains 1.07 million question-answer pairs with human-labeled categories and user-generated question descriptions. Based on such rich information, MATINF is applicable for three major NLP tasks, including classification, question answering, and summarization. We benchmark existing methods and a novel multi-task baseline over MATINF to inspire further research. Our comprehensive comparison and experiments over MATINF and other datasets demonstrate the merits held by MATINF.

MIND: A Large-scale Dataset for News Recommendation [Website][PDF]
Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, and Ming Zhou 12:00–13:00

News recommendation is an important technique for personalized news service. Compared with product and movie recommendations which have been comprehensively studied, the research on news recommendation is much more limited, mainly due to the lack of a high-quality benchmark dataset. In this paper, we present a large-scale dataset named MIND for news recommendation. Constructed from the user click logs of Microsoft News, MIND contains 1 million users and more than 160k English news articles, each of which has rich textual content such as title, abstract and body. We demonstrate MIND a good testbed for news recommendation through a comparative study of several state-of-the-art news recommendation methods which are originally developed on different proprietary datasets. Our results show the performance of news recommendation highly relies on the quality of news content understanding and user interest modeling. Many natural language processing techniques such as effective text representation methods and pre-trained language models can effectively improve the performance of news recommendation. The MIND dataset will be available at https://msnews.github.io.

ParaCrawl: Web-Scale Acquisition of Parallel Corpora [Website][PDF]
Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza 12:00–13:00

We report on methods to create the largest publicly available parallel corpora by crawling the web, using open source software. We empirically compare alternative methods and publish benchmark data sets for sentence alignment and sentence pair filtering. We also describe the parallel corpora released and evaluate their quality and their usefulness to create machine translation systems.

Session 8A Semantics: Lexical-6

Adaptive Compression of Word Embeddings [Website][PDF]
Yeachan Kim, Kang-Min Kim, and SangKeun Lee 12:00–13:00

Distributed representations of words have been an indispensable component for natural language processing (NLP) tasks. However, the large memory footprint of word embeddings makes it challenging to deploy NLP models to memory-constrained devices (e.g., self-driving cars, mobile devices). In this paper, we propose a novel method to adaptively compress word embeddings. We fundamentally follow a code-book approach that represents words as discrete codes such as (8, 5, 2, 4). However, unlike prior works that assign the same length of codes to all words, we adaptively assign different lengths of codes to each word by learning downstream tasks. The proposed method works in two steps. First, each word directly learns to select its code length in an end-to-end manner by applying the Gumbel-softmax tricks. After selecting the code length, each word learns discrete codes through a neural network with a binary constraint. To showcase the general applicability of the proposed method, we evaluate the performance on four different downstream tasks. Comprehensive evaluation results clearly show that our method is effective and makes the highly compressed word embeddings without hurting the task accuracy. Moreover, we show that our model assigns word to each code-book by considering the significance of tasks.

Analysing Lexical Semantic Change with Contextualised Word Representations [Website][PDF]
Mario Giulianelli, Marco Del Tredici, and Raquel Fernández 12:00–13:00

This paper presents the first unsupervised approach to lexical semantic change that makes use of contextualised word representations. We propose a novel method that exploits the BERT neural language model to obtain representations of word usages, clusters these representations into usage types, and measures change along time with three proposed metrics. We create a new evaluation dataset and show that the model representations and the detected semantic shifts are positively correlated with human judgements. Our extensive qualitative analysis demonstrates that our method captures a variety of synchronic and diachronic linguistic phenomena. We expect our work to inspire further research in this direction.

Autoencoding Keyword Correlation Graph for Document Clustering [Website][PDF]
Billy Chiu, Sunil Kumar Sahu, Derek Thomas, Neha Sengupta, and Mohammady Mahdy 12:00–13:00

Document clustering requires a deep understanding of the complex structure of long-text; in particular, the intra-sentential (local) and inter-sentential features (global). Existing representation learning models do not fully capture these features. To address this, we present a novel graph-based representation for document clustering that builds a *graph autoencoder* (GAE) on a Keyword Correlation Graph. The graph is constructed with topical keywords as nodes and multiple local and global features as edges. A GAE is employed to aggregate the two sets of features by learning a latent representation which can jointly reconstruct them. Clustering is then performed on the learned representations, using vector dimensions as features for inducing document classes. Extensive experiments on two datasets show that the features learned by our approach can achieve better clustering performance than other existing features, including term frequency-inverse document frequency and average embedding.

Autoencoding Pixies: Amortised Variational Inference with Graph Convolutions for Functional Distributional Semantics [Website][PDF]
Guy Emerson 12:00–13:00

Functional Distributional Semantics provides a linguistically interpretable framework for distributional semantics, by representing the meaning of a word as a function (a binary classifier), instead of a vector. However, the large number of latent variables means that inference is computationally expensive, and training a model is therefore slow to converge. In this paper, I introduce the Pixie Autoencoder, which augments the generative model of Functional Distributional Semantics with a graph-convolutional neural network to perform amortised variational inference. This allows the model to be trained more effectively, achieving better results on two tasks (semantic similarity in context and semantic composition), and outperforming BERT, a large pre-trained language model.

BERTRAM: Improved Word Embeddings Have Big Impact on Contextualized Model Performance [Website][PDF]
Timo Schick and Hinrich Schütze 12:00–13:00

Pretraining deep language models has led to large performance gains in NLP. Despite this success, Schick and Schütze (2020) recently showed that these models struggle to understand rare words. For static word embeddings, this problem has been addressed by separately learning representations for rare words. In this work, we transfer this idea to pretrained language models: We introduce BERTRAM, a powerful architecture based on BERT that is capable of inferring high-quality embeddings for rare words that are suitable as input representations for deep language models. This is achieved by enabling the surface form and contexts of a word to interact with each other in a deep architecture. Integrating BERTRAM into BERT leads to large performance increases due to improved representations of rare and medium frequency words on both a rare word probing task and three downstream tasks.

CluBERT: A Cluster-Based Approach for Learning Sense Distributions in Multiple Languages [Website][PDF]
Tommaso Pasini, Federico Scozzafava, and Bianca Scarlini 12:00–13:00

Knowing the Most Frequent Sense (MFS) of a word has been proved to help Word Sense Disambiguation (WSD) models significantly. However, the scarcity of sense-annotated data makes it difficult to induce a reliable and high-coverage distribution of the meanings in a language vocabulary. To address this issue, in this paper we present Clu-

BERT, an automatic and multilingual approach for inducing the distributions of word senses from a corpus of raw sentences. Our experiments show that CluBERT learns distributions over English senses that are of higher quality than those extracted by alternative approaches. When used to induce the MFS of a lemma, CluBERT attains state-of-the-art results on the English Word Sense Disambiguation tasks and helps to improve the disambiguation performance of two off-the-shelf WSD models. Moreover, our distributions also prove to be effective in other languages, beating all their alternatives for computing the MFS on the multilingual WSD tasks. We release our sense distributions in five different languages at https://github.com/SapienzaNLP/clubert.

Hypernymy Detection for Low-Resource Languages via Meta Learning [Website][PDF]
Changlong Yu, Jialong Han, Haisong Zhang, and Wilfred Ng 12:00–13:00

Hypernymy detection, a.k.a, lexical entailment, is a fundamental sub-task of many natural language understanding tasks. Previous explorations mostly focus on monolingual hypernymy detection on high-resource languages, e.g., English, but few investigate the low-resource scenarios. This paper addresses the problem of low-resource hypernymy detection by combining high-resource languages. We extensively compare three joint training paradigms and for the first time propose applying meta learning to relieve the low-resource issue. Experiments demonstrate the superiority of our method among the three settings, which substantially improves the performance of extremely low-resource languages by preventing over-fitting on small datasets.

Investigating Word-Class Distributions in Word Vector Spaces [Website][PDF]
Ryohei Sasano and Anna Korhonen 12:00–13:00

This paper presents an investigation on the distribution of word vectors belonging to a certain word class in a pre-trained word vector space. To this end, we made several assumptions about the distribution, modeled the distribution accordingly, and validated each assumption by comparing the goodness of each model. Specifically, we considered two types of word classes — the semantic class of direct objects of a verb and the semantic class in a thesaurus — and tried to build models that properly estimate how likely it is that a word in the vector space is a member of a given word class. Our results on selectional preference and WordNet datasets show that the centroid-based model will fail to achieve good enough performance, the geometry of the distribution and the existence of subgroups will have limited impact, and also the negative instances need to be considered for adequate modeling of the distribution. We further investigated the relationship between the scores calculated by each model and the degree of membership and found that discriminative learning-based models are best in finding the boundaries of a class, while models based on the offset between positive and negative instances perform best in determining the degree of membership.

Session 8A Semantics: Sentence Level-4

AMR Parsing with Latent Structural Information [Website][PDF]
Qiji Zhou, Yue Zhang, Donghong Ji, and Hao Tang 12:00–13:00

Abstract Meaning Representations (AMRs) capture sentence-level semantics structural representations to broad-coverage natural sentences. We investigate parsing AMR with explicit dependency structures and interpretable latent structures. We generate the latent soft structure without additional annotations, and fuse both dependency and latent structure via an extended graph neural networks. The fused structural information helps our experiments results to achieve the best reported results on both AMR 2.0 (77.5% Smatch F1 on LDC2017T10) and AMR 1.0 ((71.8% Smatch F1 on LDC2014T12).

TaPas: Weakly Supervised Table Parsing via Pre-training [Website][PDF]
Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos
12:00–13:00

Answering natural language questions over tables is usually seen as a semantic parsing task. To alleviate the collection cost of full logical forms, one popular approach focuses on weak supervision consisting of denotations instead of logical forms. However, training semantic parsers from weak supervision poses difficulties, and in addition, the generated logical forms are only used as an intermediate step prior to retrieving the denotation. In this paper, we present TaPas, an approach to question answering over tables without generating logical forms. TaPas trains from weak supervision, and predicts the denotation by selecting table cells and optionally applying a corresponding aggregation operator to such selection. TaPas extends BERT's architecture to encode tables as input, initializes from an effective joint pre-training of text segments and tables crawled from Wikipedia, and is trained end-to-end. We experiment with three different semantic parsing datasets, and find that TaPas outperforms or rivals semantic parsing models by improving state-of-the-art accuracy on SQA from 55.1 to 67.2 and performing on par with the state-of-the-art on WikiSQL and WikiTQ, but with a simpler model architecture. We additionally find that transfer learning, which is trivial in our setting, from WikiSQL to WikiTQ, yields 48.7 accuracy, 4.2 points above the state-of-the-art.

Session 8A: Sentiment Analysis, Stylistic Analysis, and Argument Mining-5

Adversarial and Domain-Aware BERT for Cross-Domain Sentiment Analysis [Website][PDF]
Chunning Du, Haifeng Sun, Jingyu Wang, Qi Qi, and Jianxin Liao 12:00–13:00

Cross-domain sentiment classification aims to address the lack of massive amounts of labeled data. It demands to predict sentiment polarity on a target domain utilizing a classifier learned from a source domain. In this paper, we investigate how to efficiently apply the pre-training language model BERT on the unsupervised domain adaptation. Due to the pre-training task and corpus, BERT is task-agnostic, which lacks domain awareness and can not distinguish the characteristic of source and target domain when transferring knowledge. To tackle these problems, we design a post-training procedure, which contains the target domain masked language model task and a novel domain-distinguish pre-training task. The post-training procedure will encourage BERT to be domain-aware and distill the domain-specific features in a self-supervised way. Based on this, we could then conduct the adversarial training to derive the enhanced domain-invariant features. Extensive experiments on Amazon dataset show that our model outperforms state-of-the-art methods by a large margin. The ablation study demonstrates that the remarkable improvement is not only from BERT but also from our method.

Don't Eclipse Your Arts Due to Small Discrepancies: Boundary Repositioning with a Pointer Network for Aspect Extraction [Website][PDF]
Zhenkai Wei, Yu Hong, Bowei Zou, Meng Cheng, and Jianmin YAO 12:00–13:00

The current aspect extraction methods suffer from boundary errors. In general, these errors lead to a relatively minor difference between the extracted aspects and the ground-truth. However, they hurt the performance severely. In this paper, we propose to utilize a pointer network for repositioning the boundaries. Recycling mechanism is used, which enables the training data to be collected without manual intervention. We conduct the experiments on the benchmark datasets SE14 of laptop and SE14-16 of restaurant. Experimental results show that our method achieves substantial improvements over the baseline, and outperforms state-of-the-art methods.

Relational Graph Attention Network for Aspect-based Sentiment Analysis [Website][PDF]
Kai Wang, Weizhou Shen, Yunyi Yang, Xiaojun Quan, and Rui Wang 12:00–13:00

Aspect-based sentiment analysis aims to determine the sentiment polarity towards a specific aspect in online reviews. Most recent efforts adopt attention-based neural network models to implicitly connect aspects with opinion words. However, due to the complexity of language and the existence of multiple aspects in a single sentence, these models often confuse the connections. In this paper, we address this problem by means of effective encoding of syntax information. Firstly, we define a unified aspect-oriented dependency tree structure rooted at a target aspect by reshaping and pruning an ordinary dependency parse tree. Then, we propose a relational graph attention network (R-GAT) to encode the new tree structure for sentiment prediction. Extensive experiments are conducted on the SemEval 2014 and Twitter datasets, and the experimental results confirm that the connections between aspects and opinion words can be better established with our approach, and the performance of the graph attention network (GAT) is significantly improved as a consequence.

SentiBERT: A Transferable Transformer-Based Architecture for Compositional Sentiment Semantics [Website][PDF]
Da Yin, Tao Meng, and Kai-Wei Chang 12:00–13:00

We propose SentiBERT, a variant of BERT that effectively captures compositional sentiment semantics. The model incorporates contextualized representation with binary constituency parse tree to capture semantic composition. Comprehensive experiments demonstrate that SentiBERT achieves competitive performance on phrase-level sentiment classification. We further demonstrate that the sentiment composition learned from the phrase-level annotations on SST can be transferred to other sentiment analysis tasks as well as related tasks, such as emotion classification tasks. Moreover, we conduct ablation studies and design visualization methods to understand SentiBERT. We show that SentiBERT is better than baseline approaches in capturing negation and the contrastive relation and model the compositional sentiment semantics.

Target Inference in Argument Conclusion Generation [Website][PDF]
Milad Alshomary, Shahbaz Syed, Martin Potthast, and Henning Wachsmuth 12:00–13:00

In argumentation, people state premises to reason towards a conclusion. The conclusion conveys a stance towards some target, such as a concept or statement. Often, the conclusion remains implicit, though, since it is self-evident in a discussion or left out for rhetorical reasons. However, the conclusion is key to understanding an argument and, hence, to any application that processes argumentation. We thus study the question to what extent an argument's conclusion can be reconstructed from its premises. In particular, we argue here that a decisive step is to infer a conclusion's target, and we hypothesize that this target is related to the premises' targets. We develop two complementary target inference approaches: one ranks premise targets and selects the top-ranked target as the conclusion target, the other finds a new conclusion target in a learned embedding space using a triplet neural network. Our evaluation on corpora from two domains indicates that a hybrid of both approaches is best, outperforming several strong baselines. According to human annotators, we infer a reasonably adequate conclusion target in 89% of the cases.

Session 8A: Student Research Workshop

How much complexity does an RNN architecture need to learn syntax-sensitive dependencies? [Website][PDF]
Gantavya Bhatt, Hritik Bansal, Rishubh Singh, and Sumeet Agarwal 12:00–13:00

Long short-term memory (LSTM) networks and their variants are capable of encapsulating long-range dependencies, which is evident from their performance on a variety of linguistic tasks. On the other hand, simple recurrent networks (SRNs), which appear more biologically grounded in terms of synaptic connections, have generally been less successful at capturing long-range dependencies as well as the loci of grammatical errors in an unsupervised setting. In this paper, we seek to develop models that bridge the gap between biological plausibility and linguistic competence. We propose a new architecture, the Decay RNN, which incorporates the decaying nature of neuronal activations and models the excitatory and inhibitory connections in a population of neurons. Besides its biological inspiration, our model also shows competitive performance relative to LSTMs on subject-verb agreement, sentence grammaticality, and language modeling tasks. These results provide some pointers towards probing the nature of the inductive biases required for RNN architectures to model linguistic phenomena successfully.

Logical Inferences with Comparatives and Generalized Quantifiers [Website][PDF]
Izumi Haruta, Koji Mineshima, and Daisuke Bekki 12:00–13:00

Comparative constructions pose a challenge in Natural Language Inference (NLI), which is the task of determining whether a text entails a hypothesis. Comparatives are structurally complex in that they interact with other linguistic phenomena such as quantifiers, numerals, and lexical antonyms. In formal semantics, there is a rich body of work on comparatives and gradable expressions using the notion of degree. However, a logical inference system for comparatives has not been sufficiently developed for use in the NLI task. In this paper, we present a compositional semantics that maps various comparative constructions in English to semantic representations via Combinatory Categorial Grammar (CCG) parsers and combine it with an inference system based on automated theorem proving. We evaluate our system on three NLI datasets that contain complex logical inferences with comparatives, generalized quantifiers, and numerals. We show that the system outperforms previous logic-based systems as well as recent deep learning-based models.

Unsupervised Multilingual Sentence Embeddings for Parallel Corpus Mining [Website][PDF]
Ivana Kvapilíková, Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Ondřej Bojar 12:00–13:00

Existing models of multilingual sentence embeddings require large parallel data resources which are not available for low-resource languages. We propose a novel unsupervised method to derive multilingual sentence embeddings relying only on monolingual data. We first produce a synthetic parallel corpus using unsupervised machine translation, and use it to fine-tune a pretrained cross-lingual masked language model (XLM) to derive the multilingual sentence representations. The quality of the representations is evaluated on two parallel corpus mining tasks with improvements of up to 22 F1 points over vanilla XLM. In addition, we observe that a single synthetic bilingual corpus is able to improve results for other language pairs.

Enhancing Word Embeddings with Knowledge Extracted from Lexical Resources [Website][PDF]
Magdalena Biesialska, Bardia Rafieian, and Marta R. Costa-jussà 12:00–13:00

In this work, we present an effective method for semantic specialization of word vector representations. To this end, we use traditional word embeddings and apply specialization methods to better capture semantic relations between words. In our approach, we leverage external knowledge from rich lexical resources such as BabelNet. We also show that our proposed post-specialization method based on an adversarial neural network with the Wasserstein distance allows to gain improvements over state-of-the-art methods on two tasks: word similarity and dialog state tracking.

Session 8A Syntax: Tagging, Chunking and Parsing-3

A Span-based Linearization for Constituent Trees [Website][PDF]
Yang Wei, Yuanbin Wu, and Man Lan 12:00–13:00

We propose a novel linearization of a constituent tree, together with a new locally normalized model. For each split point in a sentence, our model computes the normalizer on all spans ending with that split point, and then predicts a tree span from them. Compared with global models, our model is fast and parallelizable. Different from previous local models, our linearization method is tied on the spans directly and considers more local features when performing span prediction, which is more interpretable and effective. Experiments on PTB (95.8 F1) and CTB (92.4 F1) show that our model significantly outperforms existing local models and efficiently achieves competitive results with global models.

[CL] Abstract Syntax as Interlingua: Scaling Up the Grammatical Framework from Controlled Languages to Robust Pipelines [Website][PDF]
Aarne Ranta, Krasimir Angelov, Normunds Gruzitis, and Prasanth Kolachina 12:00–13:00

Abstract syntax is an interlingual representation used in compilers. Grammatical Framework (GF) applies the abstract syntax idea to natural languages. The development of GF started in 1998, first as a tool for controlled language implementations, where it has gained an established position in both academic and commercial projects. GF provides grammar resources for over 40 languages, enabling accurate generation and translation, as well as grammar engineering tools and components for mobile and Web applications. On the research side, the focus in the last ten years has been on scaling up GF to wide-coverage language processing. The concept of abstract syntax offers a unified view on many other approaches: Universal Dependencies, WordNets, FrameNets, Construction Grammars, and Abstract Meaning Representations. This makes it possible for GF to utilize data from the other approaches and to build robust pipelines. In return, GF can contribute to data-driven approaches by methods to transfer resources from one language to others, to augment data by rule-based generation, to check the consistency of hand-annotated corpora, and to pipe analyses into high-precision semantic back ends. This article gives an overview of the use of abstract syntax as interlingua through both established and emerging NLP applications involving GF.

Do Neural Language Models Show Preferences for Syntactic Formalisms? [Website][PDF]
Artur Kulmizev, Vinit Ravishankar, Mostafa Abdou, and Joakim Nivre 12:00–13:00

Recent work on the interpretability of deep neural language models has concluded that many properties of natural language syntax are encoded in their representational spaces. However, such studies often suffer from limited scope by focusing on a single language and a single linguistic formalism. In this study, we aim to investigate the extent to which the semblance of syntactic structure captured by language models adheres to a surface-syntactic or deep syntactic style of analysis, and whether the patterns are consistent across different languages. We apply a probe for extracting directed dependency trees to BERT and ELMo models trained on 13 different languages, probing for two different syntactic annotation styles: Universal Dependencies (UD), prioritizing deep syntactic relations, and Surface-Syntactic Universal Dependencies (SUD), focusing on surface structure. We find that both models exhibit a preference for UD over SUD — with interesting variations across languages and layers — and that the strength of this preference is correlated with differences in tree shape.

Enriched In-Order Linearization for Faster Sequence-to-Sequence Constituent Parsing [Website][PDF]
Daniel Fernández-González and Carlos Gómez-Rodríguez 12:00–13:00

Sequence-to-sequence constituent parsing requires a linearization to represent trees as sequences. Top-down tree linearizations, which can be based on brackets or shift-reduce actions, have achieved the best accuracy to date. In this paper, we show that these results can be improved by using an in-order linearization instead. Based on this observation, we implement an enriched in-order shift-reduce linearization inspired by Vinyals et al. (2015)'s approach, achieving the best accuracy to date on the English PTB dataset among fully-supervised single-model sequence-to-sequence constituent parsers. Finally, we apply deterministic attention mechanisms to match the speed of state-of-the-art transition-based parsers, thus showing that sequence-to-sequence models can match them, not only in accuracy, but also in speed.

Exact yet Efficient Graph Parsing, Bi-directional Locality and the Constructivist Hypothesis [Website][PDF]
Yajie Ye and Weiwei Sun 12:00–13:00

A key problem in processing graph-based meaning representations is graph parsing, i.e. computing all possible derivations of a given graph according to a (competence) grammar. We demonstrate, for the first time, that exact graph parsing can be efficient for large graphs and with large Hyperedge Replacement Grammars (HRGs). The advance is achieved by exploiting locality as terminal edge-adjacency in HRG rules. In particular, we highlight the importance of 1) a terminal edge-first parsing strategy, 2) a categorization of a subclass of HRG, i.e. what we call Weakly Regular Graph Grammar, and 3) distributing argument-structures to both lexical and phrasal rules.

Max-Margin Incremental CCG Parsing [Website][PDF]
Miloš Stanojević and Mark Steedman 12:00–13:00

Incremental syntactic parsing has been an active research area both for cognitive scientists trying to model human sentence processing and for NLP researchers attempting to combine incremental parsing with language modelling for ASR and MT. Most effort has been directed at designing the right transition mechanism, but less has been done

to answer the question of what a probabilistic model for those transition parsers should look like. A very incremental transition mechanism of a recently proposed CCG parser when trained in straightforward locally normalised discriminative fashion produces very bad results on English CCGbank. We identify three biases as the causes of this problem: label bias, exposure bias and imbalanced probabilities bias. While known techniques for tackling these biases improve results, they still do not make the parser state of the art. Instead, we tackle all of these three biases at the same time using an improved version of beam search optimisation that minimises all beam search violations instead of minimising only the biggest violation. The new incremental parser gives better results than all previously published incremental CCG parsers, and outperforms even some widely used non-incremental CCG parsers.

Neural Reranking for Dependency Parsing: An Evaluation [Website] [PDF]
Bich-Ngoc Do and Ines Rehbein 12:00–13:00

Recent work has shown that neural rerankers can improve results for dependency parsing over the top k trees produced by a base parser. However, all neural rerankers so far have been evaluated on English and Chinese only, both languages with a configurational word order and poor morphology. In the paper, we re-assess the potential of successful neural reranking models from the literature on English and on two morphologically rich(er) languages, German and Czech. In addition, we introduce a new variation of a discriminative reranker based on graph convolutional networks (GCNs). We show that the GCN not only outperforms previous models on English but is the only model that is able to improve results over the baselines on German and Czech. We explain the differences in reranking performance based on an analysis of a) the gold tree ratio and b) the variety in the k-best lists.

Demo Session 3B

Time: *12:45–13:30*

CLIReval: Evaluating Machine Translation as a Cross-Lingual Information Retrieval Task [Website][PDF]

Shuo Sun, Suzanna Sia, and Kevin Duh

We present CLIReval, an easy-to-use toolkit for evaluating machine translation (MT) with the proxy task of cross-lingual information retrieval (CLIR). Contrary to what the project name might suggest, CLIReval does not actually require any annotated CLIR dataset. Instead, it automatically transforms translations and references used in MT evaluations into a synthetic CLIR dataset; it then sets up a standard search engine (Elasticsearch) and computes various information retrieval metrics (e.g., mean average precision) by treating the translations as documents to be retrieved. The idea is to gauge the quality of MT by its impact on the document translation approach to CLIR. As a case study, we run CLIReval on the "metrics shared task" of WMT2019; while this extrinsic metric is not intended to replace popular intrinsic metrics such as BLEU, results suggest CLIReval is competitive in many language pairs in terms of correlation to human judgments of quality. CLIReval is publicly available at https://github.com/ssun32/CLIReval.

Label Noise in Context [Website][PDF]

Michael Desmond, Catherine Finegan-Dollak, Jeff Boston, and Matt Arnold

Label noise—incorrectly or ambiguously labeled training examples—can negatively impact model performance. Although noise detection techniques have been around for decades, practitioners rarely apply them, as manual noise remediation is a tedious process. Examples incorrectly flagged as noise waste reviewers' time, and correcting label noise without guidance can be difficult. We propose LNIC, a noise-detection method that uses an example's neighborhood within the training set to (a) reduce false positives and (b) provide an explanation as to why the ex- ample was flagged as noise. We demonstrate on several short-text classification datasets that LNIC outperforms the state of the art on measures of precision and F0.5-score. We also show how LNIC's training set context helps a reviewer to understand and correct label noise in a dataset. The LNIC tool lowers the barriers to label noise remediation, increasing its utility for NLP practitioners.

Session 8B Overview – Tuesday, July 7, 2020 13:00–14:00

Track A *Ethics and NLP-3* Abstracts	Demographics Should Not Be the Reason of Toxicity: Mitigating Discrimination in Text Classifications with Instance Weighting *Zhang, Bai, Zhang, Bai, Zhu, and Zhao* [Website][PDF]	It's Morphin' Time! Combating Linguistic Discrimination with Inflectional Perturbations *Tan, Joty, Kan, and Socher* [Website][PDF]	Toward Gender-Inclusive Coreference Resolution *Cao and Daumé III* [Website][PDF]		
Track B *Generation-10* Abstracts	Learning Implicit Text Generation via Feature Matching *Padhi, Dognin, Bai, Nogueira dos Santos, Chenthamarakshan, Mroueh, and Das* [Website][PDF]	Two Birds, One Stone: A Simple, Unified Model for Text Generation from Structured and Unstructured Data *Shahidi, Li, and Lin* [Website][PDF]			
Track C *Interpretability and Analysis of Models for NLP-4* Abstracts	[TACL] Does Syntax Need to Grow on Trees? Sources of Hierarchical Inductive Bias in Sequence-to-Sequence Networks *McCoy, Frank, and Linzen* [Website][PDF]	Human Attention Maps for Text Classification: Do Humans and Neural Networks Focus on the Same Words? *Sen, Hartvigsen, Yin, Kong, and Rundensteiner* [Website][PDF]	Information-Theoretic Probing for Linguistic Structure *Pimentel, Valvoda, Hall Maudslay, Zmigrod, Williams, and Cotterell* [Website][PDF]	[TACL] Membership Inference Attacks on Sequence-to-Sequence Models: Is My Data In Your Machine Translation System? *Hisamoto, Post, and Duh* [Website][PDF]	On the Cross-lingual Transferability of Monolingual Representations *Artetxe, Ruder, and Yogatama* [Website][PDF]
	Perturbed Masking: Parameter-free Probing for Analyzing and Interpreting BERT *Wu, Chen, Kao, and Liu* [Website][PDF]	Quantifying Attention Flow in Transformers *Abnar and Zuidema* [Website][PDF]	Similarity Analysis of Contextual Word Representation Models *Wu, Belinkov, Sajjad, Durrani, Dalvi, and Glass* [Website][PDF]	[TACL] Tabula nearly rasa: Probing the linguistic knowledge of character-level neural language models trained on unsegmented text *Hahn and Baroni* [Website][PDF]	Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? *Jacovi and Goldberg* [Website][PDF]
	[TACL] What BERT is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models *Ettinger* [Website][PDF]				
Track D *Machine Learning for NLP-9* Abstracts	Attentive Pooling with Learnable Norms for Text Representation *Wu, Wu, Qi, Cui, and Huang* [Website][PDF]	Bayesian Hierarchical Words Representation Learning *Barkan, Rejwan, Caciularu, and Koenigstein* [Website][PDF]	On the Encoder-Decoder Incompatibility in Variational Text Modeling and Beyond *Wu, Wang, and Wang* [Website][PDF]	Pre-training Is (Almost) All You Need: An Application to Commonsense Reasoning *Tamborrino, Pellicanò, Pannier, Voitot, and Naudin* [Website][PDF]	SEEK: Segmented Embedding of Knowledge Graphs *Xu, Zheng, He, Shao, Yin, and Liu* [Website][PDF]

	SenseBERT: Driving Some Sense into BERT *Levine, Lenz, Dagan, Ram, Padnos, Sharir, Shalev-Shwartz, Shashua, and Shoham* [Website][PDF]	Single Model Ensemble using Pseudo-Tags and Distinct Vectors *Kuwabara, Suzuki, and Nakayama* [Website][PDF]	Tchebycheff Procedure for Multi-task Text Classification *Mao, Yun, Liu, and Du* [Website][PDF]		
Track E *Machine Translation-12* Abstracts	A Relaxed Matching Procedure for Unsupervised BLI *Zhao, Wang, Zhang, and Wu* [Website][PDF]	[CL] A Systematic Study of Inner-Attention-Based Sentence Representations in Multilingual Neural Machine Translation *Vázquez, Raganato, Creutz, and Tiedemann* [Website][PDF]	[TACL] Better Document-level Machine Translation with Bayes' Rule *Yu, Sartran, Stokowiec, Ling, Kong, Blunsom, and Dyer* [Website]	Selecting Back-translated Data from Multiple Sources for Improved Neural Machine Translation *Soto, Shterionov, Poncelas, and Way* [Website][PDF]	[CL] Unsupervised Word Translation with Adversarial Autoencoder *Mohiuddin and Joty* [Website][PDF]
Track F *NLP Applications-7* Abstracts	Empowering Active Learning to Jointly Optimize System and User Demands *Lee, Meyer, and Gurevych* [Website][PDF]	Encoder-Decoder Models Can Benefit from Pre-trained Masked Language Models in Grammatical Error Correction *Kaneko, Mita, Kiyono, Suzuki, and Inui* [Website][PDF]	Graph Neural News Recommendation with Unsupervised Preference Disentanglement *Hu, Xu, Li, Yang, Shi, Duan, Xie, and Zhou* [Website][PDF]	Hiring Now: A Skill-Aware Multi-Attention Model for Job Posting Generation *Liu, Liu, Zhang, Chi, Shi, and Huang* [Website][PDF]	Identifying Principals and Accessories in a Complex Case based on the Comprehension of Fact Description *Hu, Luo, and Chao* [Website][PDF]
	Joint Modelling of Emotion and Abusive Language Detection *Rajamanickam, Mishra, Yannakoudakis, and Shutova* [Website][PDF]	Programming in Natural Language with fuSE: Synthesizing Methods from Spoken Utterances Using Deep Natural Language Understanding *Weigelt, Steurer, Hey, and Tichy* [Website][PDF]	Toxicity Detection: Does Context Really Matter? *Pavlopoulos, Sorensen, Dixon, Thain, and Androutsopoulos* [Website][PDF]		
Track G *Resources and Evaluation-8* Abstracts	ASSET: A Dataset for Tuning and Evaluation of Sentence Simplification Models with Multiple Rewriting Transformations *Alva-Manchego, Martin, Bordes, Scarton, Sagot, and Specia* [Website][PDF]	Automatic Machine Translation Evaluation using Source Language Inputs and Cross-lingual Language Model *Takahashi, Sudoh, and Nakamura* [Website][PDF]	Fatality Killed the Cat or: BabelPic, a Multimodal Dataset for Non-Concrete Concepts *Calabrese, Bevilacqua, and Navigli* [Website][PDF]	[TACL] Paraphrase-Sense-Tagged Sentences *Cocos and Callison-Burch* [Website][PDF]	That is a Known Lie: Detecting Previously Fact-Checked Claims *Shaar, Babulkov, Da San Martino, and Nakov* [Website][PDF]
Track H *Sentiment Analysis, Stylistic Analysis, and Argument Mining-6* Abstracts	Aspect Sentiment Classification with Document-level Sentiment Preference Modeling *Chen, Sun, Wang, Li, Si, Zhang, and Zhou* [Website][PDF]	ECPE-2D: Emotion-Cause Pair Extraction based on Joint Two-Dimensional Representation, Interaction and Prediction *Ding, Xia, and Yu* [Website][PDF]	From Arguments to Key Points: Towards Automatic Argument Summarization *Bar-Haim, Eden, Friedman, Kantor, Lahav, and Slonim* [Website][PDF]	He said "who's gonna take care of your children when you are at ACL?": Reported Sexist Acts are Not Sexist *Chiril, MORICEAU, Benamara, Mari, Origgi, and Coulomb-Gully* [Website][PDF]	Modeling Label Semantics for Predicting Emotional Reactions *Gaonkar, Kwon, Bastan, Balasubramanian, and Chambers* [Website][PDF]

	SKEP: Sentiment Knowledge Enhanced Pre-training for Sentiment Analysis *Tian, Gao, Xiao, Liu, He, Wu, Wang, and* [Website][PDF]				
Track I *Speech and Multimodality-5* Abstracts	How Accents Confound: Probing for Accent Information in End-to-End Speech Recognition Systems *Prasad and Jyothi* [Website][PDF]	Learning Spoken Language Representations with Neural Lattice Language Modeling *Huang and Chen* [Website][PDF]	Meta-Transfer Learning for Code-Switched Speech Recognition *Winata, Cahyawijaya, Lin, Liu, Xu, and Fung* [Website][PDF]	Multimodal Transformer for Multimodal Machine Translation *Yao and Wan* [Website][PDF]	Sentiment and Emotion help Sarcasm? A Multi-task Learning Framework for Multi-Modal Sarcasm, Sentiment and Emotion Analysis *Chauhan, S R, Ekbal, and Bhattacharyya* [Website][PDF]
	SimulSpeech: End-to-End Simultaneous Speech to Text Translation *Ren, Liu, Tan, Zhang, QIN, Zhao, and Liu* [Website][PDF]	Towards Emotion-aided Multi-modal Dialogue Act Classification *Saha, Patra, Saha, and Bhattacharyya* [Website][PDF]			
Track J *Student Research Workshop* Abstracts	Pre-training via Leveraging Assisting Languages for Neural Machine Translation *Song, Dabre, Mao, Cheng, Kurohashi, and Sumita* [Website][PDF]	Preventing Critical Scoring Errors in Short Answer Scoring with Confidence Estimation *Funayama, Sasaki, Matsubayashi, Mizumoto, Suzuki, Mita, and Inui* [Website][PDF]			

Session 8B Details

Session 8B: Ethics and NLP-3

Demographics Should Not Be the Reason of Toxicity: Mitigating Discrimination in Text Classifications with Instance Weighting [Website][PDF]
Guanhua Zhang, Bing Bai, Junqi Zhang, Kun Bai, Conghui Zhu, and Tiejun Zhao 13:00–14:00

With the recent proliferation of the use of text classifications, researchers have found that there are certain unintended biases in text classification datasets. For example, texts containing some demographic identity-terms (e.g., "gay", "black") are more likely to be abusive in existing abusive language detection datasets. As a result, models trained with these datasets may consider sentences like "She makes me happy to be gay" as abusive simply because of the word "gay." In this paper, we formalize the unintended biases in text classification datasets as a kind of selection bias from the non-discrimination distribution to the discrimination distribution. Based on this formalization, we further propose a model-agnostic debiasing training framework by recovering the non-discrimination distribution using instance weighting, which does not require any extra resources or annotations apart from a pre-defined set of demographic identity-terms. Experiments demonstrate that our method can effectively alleviate the impacts of the unintended biases without significantly hurting models' generalization ability.

It's Morphin' Time! Combating Linguistic Discrimination with Inflectional Perturbations [Website][PDF]
Samson Tan, Shafiq Joty, Min-Yen Kan, and Richard Socher 13:00–14:00

Training on only perfect Standard English corpora predisposes pre-trained neural networks to discriminate against minorities from non-standard linguistic backgrounds (e.g., African American Vernacular English, Colloquial Singapore English, etc.). We perturb the inflectional morphology of words to craft plausible and semantically similar adversarial examples that expose these biases in popular NLP models, e.g., BERT and Transformer, and show that adversarially fine-tuning them for a single epoch significantly improves robustness without sacrificing performance on clean data.

Toward Gender-Inclusive Coreference Resolution [Website][PDF]
Yang Trista Cao and Hal Daumé III 13:00–14:00

Correctly resolving textual mentions of people fundamentally entails making inferences about those people. Such inferences raise the risk of systemic biases in coreference resolution systems, including biases that can harm binary and non-binary trans and cis stakeholders. To better understand such biases, we foreground nuanced conceptualizations of gender from sociology and sociolinguistics, and develop two new datasets for interrogating bias in crowd annotations and in existing coreference resolution systems. Through these studies, conducted on English text, we confirm that without acknowledging and building systems that recognize the complexity of gender, we build systems that lead to many potential harms.

Session 8B: Generation-10

Learning Implicit Text Generation via Feature Matching [Website][PDF]
Inkit Padhi, Pierre Dognin, Ke Bai, Cícero Nogueira dos Santos, Vijil Chenthamarakshan, Youssef Mroueh, and Payel Das 13:00–14:00

Generative feature matching network (GFMN) is an approach for training state-of-the-art implicit generative models for images by performing moment matching on features from pre-trained neural networks. In this paper, we present new GFMN formulations that are effective for sequential data. Our experimental results show the effectiveness of the proposed method, SeqGFMN, for three distinct generation tasks in English: unconditional text generation, class-conditional text generation, and unsupervised text style transfer. SeqGFMN is stable to train and outperforms various adversarial approaches for text generation and text style transfer.

Two Birds, One Stone: A Simple, Unified Model for Text Generation from Structured and Unstructured Data [Website][PDF]
Hamidreza Shahidi, Ming Li, and Jimmy Lin 13:00–14:00

A number of researchers have recently questioned the necessity of increasingly complex neural network (NN) architectures. In particular, several recent papers have shown that simpler, properly tuned models are at least competitive across several NLP tasks. In this work, we show that this is also the case for text generation from structured and unstructured data. We consider neural table-to-text generation and neural question generation (NQG) tasks for text generation from structured and unstructured data, respectively. Table-to-text generation aims to generate a description based on a given table, and NQG is the task of generating a question from a given passage where the generated question can be answered by a certain sub-span of the passage using NN models. Experimental results demonstrate that a basic attention-based seq2seq model trained with the exponential moving average technique achieves the state of the art in both tasks. Code is available at https://github.com/h-shahidi/2birds-gen.

Session 8B: Interpretability and Analysis of Models for NLP-4

[TACL] Does Syntax Need to Grow on Trees? Sources of Hierarchical Inductive Bias in Sequence-to-Sequence Networks [Website][PDF]
R. Thomas McCoy, Robert Frank, and Tal Linzen 13:00–14:00

Learners that are exposed to the same training data might generalize differently due to differing inductive biases. In neural network models, inductive biases could in theory arise from any aspect of the model architecture. We investigate which architectural factors affect the generalization behavior of neural sequence-to-sequence models trained on two syntactic tasks, English question formation and English tense reinflection. For both tasks, the training set is consistent with a generalization based on hierarchical structure and a generalization based on linear order. All architectural factors that we investigated qualitatively affected how models generalized, including factors with no clear connection to hierarchical structure. For example, LSTMs and GRUs displayed qualitatively different inductive biases. However, the only factor that consistently contributed a hierarchical bias across tasks was the use of a tree-structured model rather than a model with sequential recurrence, suggesting that human-like syntactic generalization requires architectural syntactic structure.

Human Attention Maps for Text Classification: Do Humans and Neural Networks Focus on the Same Words? [Website][PDF]
Cansu Sen, Thomas Hartvigsen, Biao Yin, Xiangnan Kong, and Elke Rundensteiner 13:00–14:00

Motivated by human attention, computational attention mechanisms have been designed to help neural networks adjust their focus on specific parts of the input data. While attention mechanisms are claimed to achieve interpretability, little is known about the actual relationships between machine and human attention. In this work, we conduct the first quantitative assessment of human versus computational attention mechanisms for the text classification task. To achieve this, we design and conduct a large-scale crowd-sourcing study to collect human attention maps that encode the parts of a text that humans focus on when conducting text classification. Based on this new resource of human attention dataset for text classification, YELP-HAT, collected on the publicly available YELP dataset, we perform a quantitative comparative analysis of machine attention maps created by deep learning models and human attention maps. Our analysis offers insights into the relationships between human versus machine attention maps along three dimensions: overlap in word selections, distribution over lexical categories, and context-dependency of sentiment polarity. Our findings open promising future research opportunities ranging from supervised attention to the design of human-centric attention-based explanations.

Information-Theoretic Probing for Linguistic Structure [Website][PDF]
Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell
13:00–14:00

The success of neural networks on a diverse set of NLP tasks has led researchers to question how much these networks actually "know" about natural language. Probes are a natural way of assessing this. When probing, a researcher chooses a linguistic task and trains a supervised model to predict annotations in that linguistic task from the network's learned representations. If the probe does well, the researcher may conclude that the representations encode knowledge related to the task. A commonly held belief is that using simpler models as probes is better; the logic is that simpler models will identify linguistic structure, but not learn the task itself. We propose an information-theoretic operationalization of probing as estimating mutual information that contradicts this received wisdom: one should always select the highest performing probe one can, even if it is more complex, since it will result in a tighter estimate, and thus reveal more of the linguistic information inherent in the representation. The experimental portion of our paper focuses on empirically estimating the mutual information between a linguistic property and BERT, comparing these estimates to several baselines. We evaluate on a set of ten typologically diverse languages often underrepresented in NLP research—plus English—totalling eleven languages. Our implementation is available in https://github.com/rycolab/info-theoretic-probing.

[TACL] Membership Inference Attacks on Sequence-to-Sequence Models: Is My Data In Your Machine Translation System? [Website][PDF]
Sorami Hisamoto, Matt Post, and Kevin Duh 13:00–14:00

Data privacy is an important issue for "machine learning as a service" providers. We focus on the problem of membership inference attacks: given a data sample and black-box access to a model's API, determine whether the sample existed in the model's training data. Our contribution is an investigation of this problem in the context of sequence-to-sequence models, which are important in applications such as machine translation and video captioning. We define the membership inference problem for sequence generation, provide an open dataset based on state-of-the-art machine translation models, and report initial results on whether these models leak private information against several kinds of membership inference attacks.

On the Cross-lingual Transferability of Monolingual Representations [Website][PDF]
Mikel Artetxe, Sebastian Ruder, and Dani Yogatama 13:00–14:00

State-of-the-art unsupervised multilingual models (e.g., multilingual BERT) have been shown to generalize in a zero-shot cross-lingual setting. This generalization ability has been attributed to the use of a shared subword vocabulary and joint training across multiple languages giving rise to deep multilingual abstractions. We evaluate this hypothesis by designing an alternative approach that transfers a monolingual model to new languages at the lexical level. More concretely, we first train a transformer-based masked language model on one language, and transfer it to a new language by learning a new embedding matrix with the same masked language modeling objective, freezing parameters

of all other layers. This approach does not rely on a shared vocabulary or joint training. However, we show that it is competitive with multilingual BERT on standard cross-lingual classification benchmarks and on a new Cross-lingual Question Answering Dataset (XQuAD). Our results contradict common beliefs of the basis of the generalization ability of multilingual models and suggest that deep monolingual models learn some abstractions that generalize across languages. We also release XQuAD as a more comprehensive cross-lingual benchmark, which comprises 240 paragraphs and 1190 question-answer pairs from SQuAD v1.1 translated into ten languages by professional translators.

Perturbed Masking: Parameter-free Probing for Analyzing and Interpreting BERT [Website][PDF]
Zhiyong Wu, Yun Chen, Ben Kao, and Qun Liu 13:00–14:00

By introducing a small set of additional parameters, a *probe* learns to solve specific linguistic tasks (e.g., dependency parsing) in a supervised manner using feature representations (e.g., contextualized embeddings). The effectiveness of such *probing* tasks is taken as evidence that the pre-trained model encodes linguistic knowledge. However, this approach of evaluating a language model is undermined by the uncertainty of the amount of knowledge that is learned by the probe itself. Complementary to those works, we propose a parameter-free probing technique for analyzing pre-trained language models (e.g., BERT). Our method does not require direct supervision from the probing tasks, nor do we introduce additional parameters to the probing process. Our experiments on BERT show that syntactic trees recovered from BERT using our method are significantly better than linguistically-uninformed baselines. We further feed the empirically induced dependency structures into a downstream sentiment classification task and find its improvement compatible with or even superior to a human-designed dependency schema.

Quantifying Attention Flow in Transformers [Website][PDF]
Samira Abnar and Willem Zuidema 13:00–14:00

In the Transformer model, "self-attention" combines information from attended embeddings into the representation of the focal embedding in the next layer. Thus, across layers of the Transformer, information originating from different tokens gets increasingly mixed. This makes attention weights unreliable as explanations probes. In this paper, we consider the problem of quantifying this flow of information through self-attention. We propose two methods for approximating the attention to input tokens given attention weights, attention rollout and attention flow, as post hoc methods when we use attention weights as the relative relevance of the input tokens. We show that these methods give complementary views on the flow of information, and compared to raw attention, both yield higher correlations with importance scores of input tokens obtained using an ablation method and input gradients.

Similarity Analysis of Contextual Word Representation Models [Website][PDF]
John Wu, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass 13:00–14:00

This paper investigates contextual word representation models from the lens of similarity analysis. Given a collection of trained models, we measure the similarity of their internal representations and attention. Critically, these models come from vastly different architectures. We use existing and novel similarity measures that aim to gauge the level of localization of information in the deep models, and facilitate the investigation of which design factors affect model similarity, without requiring any external linguistic annotation. The analysis reveals that models within the same family are more similar to one another, as may be expected. Surprisingly, different architectures have rather similar representations, but different individual neurons. We also observed differences in information localization in lower and higher layers and found that higher layers are more affected by fine-tuning on downstream tasks.

[TACL] Tabula nearly rasa: Probing the linguistic knowledge of character-level neural language models trained on unsegmented text [Website][PDF]
Michael Hahn and Marco Baroni 13:00–14:00

Recurrent neural networks (RNNs) reached striking performance in many natural language processing tasks. This has renewed interest in whether these generic sequence processing devices are inducing genuine linguistic knowledge. Nearly all current analytical studies, however, initialize the RNNs with a vocabulary of known words, and feed them tokenized input during training. We present a multi-lingual study of the linguistic knowledge encoded in RNNs trained as character-level language models, on input data with word boundaries removed. These networks face a tougher and more cognitively realistic task, having to discover and store any useful linguistic unit from scratch, based on input statistics. The results show that our "near tabula rasa" RNNs are mostly able to solve morphological, syntactic and semantic tasks that intuitively presuppose word-level knowledge, and indeed they learned to track "soft" word boundaries. Our study opens the door to speculations about the necessity of an explicit word lexicon in language learning and usage.

Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? [Website][PDF]
Alon Jacovi and Yoav Goldberg 13:00–14:00

With the growing popularity of deep-learning based NLP models, comes a need for interpretable systems. But what is interpretability, and what constitutes a high-quality interpretation? In this opinion piece we reflect on the current state of interpretability evaluation research. We call for more clearly differentiating between different desired criteria an interpretation should satisfy, and focus on the faithfulness criteria. We survey the literature with respect to faithfulness evaluation, and arrange the current approaches around three assumptions, providing an explicit form to how faithfulness is "defined" by the community. We provide concrete guidelines on how evaluation of interpretation methods should and should not be conducted. Finally, we claim that the current binary definition for faithfulness sets a potentially unrealistic bar for being considered faithful. We call for discarding the binary notion of faithfulness in favor of a more graded one, which we believe will be of greater practical utility.

[TACL] What BERT is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models [Website] [PDF]

Allyson Ettinger 13:00–14:00

Pre-training by language modeling has become a popular and successful approach to NLP tasks, but we have yet to understand exactly what linguistic capacities these pre-training processes confer upon models. In this paper we introduce a suite of diagnostics drawn from human language experiments, which allow us to ask targeted questions about information used by language models for generating predictions in context. As a case study, we apply these diagnostics to the popular BERT model, finding that it can generally distinguish good from bad completions involving shared category or role reversal, albeit with less sensitivity than humans, and it robustly retrieves noun hypernyms, but it struggles with challenging inference and role-based event prediction – and in particular, it shows clear insensitivity to the contextual impacts of negation.

Session 8B: Machine Learning for NLP-9

Attentive Pooling with Learnable Norms for Text Representation [Website][PDF]
Chuhan Wu, Fangzhao Wu, Tao Qi, Xiaohui Cui, and Yongfeng Huang 13:00–14:00

Pooling is an important technique for learning text representations in many neural NLP models. In conventional pooling methods such as average, max and attentive pooling, text representations are weighted summations of the L1 or L∞ norm of input features. However, their pooling norms are always fixed and may not be optimal for learning accurate text representations in different tasks. In addition, in many popular pooling methods such as max and attentive pooling some features may be over-emphasized, while other useful ones are not fully exploited. In this paper, we propose an Attentive Pooling with Learnable Norms (APLN) approach for text representation. Different from existing pooling methods that use a fixed pooling norm, we propose to learn the norm in an end-to-end manner to automatically find the optimal ones for text representation in different tasks. In addition, we propose two methods to ensure the numerical stability of the model training. The first one is scale limiting, which re-scales the input to ensure non-negativity and alleviate the risk of exponential explosion. The second one is re-formulation, which decomposes the exponent operation to avoid computing the real-valued powers of the input and further accelerate the pooling operation. Experimental results on four benchmark datasets show that our approach can effectively improve the performance of attentive pooling.

Bayesian Hierarchical Words Representation Learning [Website][PDF]
Oren Barkan, Idan Rejwan, Avi Caciularu, and Noam Koenigstein 13:00–14:00

This paper presents the Bayesian Hierarchical Words Representation (BHWR) learning algorithm. BHWR facilitates Variational Bayes word representation learning combined with semantic taxonomy modeling via hierarchical priors. By propagating relevant information between related words, BHWR utilizes the taxonomy to improve the quality of such representations. Evaluation of several linguistic datasets demonstrates the advantages of BHWR over suitable alternatives that facilitate Bayesian modeling with or without semantic priors. Finally, we further show that BHWR produces better representations for rare words.

On the Encoder-Decoder Incompatibility in Variational Text Modeling and Beyond [Website][PDF]
Chen Wu, Prince Zizhuang Wang, and William Yang Wang 13:00–14:00

Variational autoencoders (VAEs) combine latent variables with amortized variational inference, whose optimization usually converges into a trivial local optimum termed posterior collapse, especially in text modeling. By tracking the optimization dynamics, we observe the encoder-decoder incompatibility that leads to poor parameterizations of the data manifold. We argue that the trivial local optimum may be avoided by improving the encoder and decoder parameterizations since the posterior network is part of a transition map between them. To this end, we propose Coupled-VAE, which couples a VAE model with a deterministic autoencoder with the same structure and improves the encoder and decoder parameterizations via encoder weight sharing and decoder signal matching. We apply the proposed Coupled-VAE approach to various VAE models with different regularization, posterior family, decoder structure, and optimization strategy. Experiments on benchmark datasets (i.e., PTB, Yelp, and Yahoo) show consistently improved results in terms of probability estimation and richness of the latent space. We also generalize our method to conditional language modeling and propose Coupled-CVAE, which largely improves the diversity of dialogue generation on the Switchboard dataset.

Pre-training Is (Almost) All You Need: An Application to Commonsense Reasoning [Website][PDF]
Alexandre Tamborrino, Nicola Pellicanò, Baptiste Pannier, Pascal Voitot, and Louise Naudin 13:00–14:00

Fine-tuning of pre-trained transformer models has become the standard approach for solving common NLP tasks. Most of the existing approaches rely on a randomly initialized classifier on top of such networks. We argue that this fine-tuning procedure is sub-optimal as the pre-trained model has no prior on the specific classifier labels, while it might have already learned an intrinsic textual representation of the task. In this paper, we introduce a new scoring method that casts a plausibility ranking task in a full-text format and leverages the masked language modeling head tuned during the pre-training phase. We study commonsense reasoning tasks where the model must rank a set of hypotheses given a premise, focusing on the COPA, Swag, HellaSwag and CommonsenseQA datasets. By exploiting our scoring method without fine-tuning, we are able to produce strong baselines (e.g. 80% test accuracy on COPA) that are comparable to supervised approaches. Moreover, when fine-tuning directly on the proposed scoring function, we show that our method provides a much more stable training phase across random restarts (e.g x10 standard deviation reduction on COPA test accuracy) and requires less annotated data than the standard classifier approach to reach equivalent performances.

SEEK: Segmented Embedding of Knowledge Graphs [Website][PDF]
Wentao Xu, Shun Zheng, Liang He, Bin Shao, Jian Yin, and Tie-Yan Liu 13:00–14:00

In recent years, knowledge graph embedding becomes a pretty hot research topic of artificial intelligence and plays increasingly vital roles in various downstream applications, such as recommendation and question answering. However, existing methods for knowledge graph embedding can not make a proper trade-off between the model complexity and the model expressiveness, which makes them still far from satisfactory. To mitigate this problem, we propose a lightweight modeling framework that can achieve highly competitive relational expressiveness without increasing the model complexity. Our framework focuses on the design of scoring functions and highlights two critical characteristics: 1) facilitating sufficient feature interactions; 2) preserving both symmetry and antisymmetry properties of relations. It is noteworthy that owing to the general and elegant design of scoring functions, our framework can incorporate many famous existing methods as special cases. Moreover, extensive experiments on public benchmarks demonstrate the efficiency and effectiveness of our framework. Source codes and data can be found at

https://github.com/Wentao-Xu/SEEK.

SenseBERT: Driving Some Sense into BERT [Website][PDF]
Yoav Levine, Barak Lenz, Or Dagan, Ori Ram, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham 13:00–14:00

The ability to learn from large unlabeled corpora has allowed neural language models to advance the frontier in natural language understanding. However, existing self-supervision techniques operate at the word form level, which serves as a surrogate for the underlying semantic content. This paper proposes a method to employ weak-supervision directly at the word sense level. Our model, named SenseBERT, is pre-trained to predict not only the masked words but also their WordNet supersenses. Accordingly, we attain a lexical-semantic level language model, without the use of human annotation. SenseBERT achieves significantly improved lexical understanding, as we demonstrate by experimenting on SemEval Word Sense Disambiguation, and by attaining a state of the art result on the 'Word in Context' task.

Single Model Ensemble using Pseudo-Tags and Distinct Vectors [Website][PDF]
Ryosuke Kuwabara, Jun Suzuki, and Hideki Nakayama 13:00–14:00

Model ensemble techniques often increase task performance in neural networks; however, they require increased time, memory, and management effort. In this study, we propose a novel method that replicates the effects of a model ensemble with a single model. Our approach creates K-virtual models within a single parameter space using K-distinct pseudo-tags and K-distinct vectors. Experiments on text classification and sequence labeling tasks on several datasets demonstrate that our method emulates or outperforms a traditional model ensemble with 1/K-times fewer parameters.

Tchebycheff Procedure for Multi-task Text Classification [Website][PDF]
Yuren Mao, Shuang Yun, Weiwei Liu, and Bo Du 13:00–14:00

Multi-task Learning methods have achieved great progress in text classification. However, existing methods assume that multi-task text classification problems are convex multiobjective optimization problems, which is unrealistic in real-world applications. To address this issue, this paper presents a novel Tchebycheff procedure to optimize the multi-task classification problems without convex assumption. The extensive experiments back up our theoretical analysis and validate the superiority of our proposals.

Session 8B: Machine Translation-12

A Relaxed Matching Procedure for Unsupervised BLI [Website][PDF]
Xu Zhao, Zihao Wang, Yong Zhang, and Hao Wu 13:00–14:00

Recently unsupervised Bilingual Lexicon Induction(BLI) without any parallel corpus has attracted much research interest. One of the crucial parts in methods for the BLI task is the matching procedure. Previous works impose a too strong constraint on the matching and lead to many counterintuitive translation pairings. Thus We propose a relaxed matching procedure to find a more precise matching between two languages. We also find that aligning source and target language embedding space bidirectionally will bring significant improvement. We follow the previous iterative framework to conduct experiments. Results on standard benchmark demonstrate the effectiveness of our proposed method, which substantially outperforms previous unsupervised methods.

[CL] A Systematic Study of Inner-Attention-Based Sentence Representations in Multilingual Neural Machine Translation [Website][PDF]
Raúl Vázquez, Alessandro Raganato, Mathias Creutz, and Jörg Tiedemann 13:00–14:00

Neural machine translation has considerably improved the quality of automatic translations by learning good representations of input sentences. In this article, we explore a multilingual translation model capable of producing fixed-size sentence representations by incorporating an intermediate crosslingual shared layer, which we refer to as attention bridge. This layer exploits the semantics from each language and develops into a language-agnostic meaning representation that can be efficiently used for transfer learning.We systematically study the impact of the size of the attention bridge and the effect of including additional languages in the model. In contrast to related previous work, we demonstrate that there is no conflict between translation performance and the use of sentence representations in downstream tasks. In particular, we show that larger intermediate layers not only improve translation quality, especially for long sentences, but also push the accuracy of trainable classification tasks. Nevertheless, shorter representations lead to increased compression that is beneficial in non-trainable similarity tasks. Similarly, we show that trainable downstream tasks benefit from multilingual models, whereas additional language signals do not improve performance in non-trainable benchmarks. This is an important insight that helps to properly design models for specific applications. Finally, we also include an in-depth analysis of the proposed attention bridge and its ability of encoding linguistic properties. We carefully analyze the information that is captured by individual attention heads and identify interesting patterns that explain the performance of specific settings in linguistic probing tasks.

[TACL] Better Document-level Machine Translation with Bayes' Rule [Website]
Lei Yu, Laurent Sartran, Wojciech Stokowiec, Wang Ling, Lingpeng Kong, Phil Blunsom, and Chris Dyer
13:00–14:00

We show that Bayes' rule provides an effective mechanism for creating document translation models that can be learned from only parallel sentences and monolingual documents—a compelling benefit as parallel documents are not always available. In our formulation, the posterior probability of a candidate translation is the product of the unconditional (prior) probability of the candidate output document and the "reverse translation probability" of translating the candidate output back into the source language. Our proposed model uses a powerful autoregressive language model as the prior on target language documents, but it assumes that each sentence is translated independently from the target to the source language. Crucially, at test time, when a source document is observed, the document language model prior induces dependencies between the translations of the source sentences in the posterior. The model's independence assumption not only enables efficient use of available data, but it additionally admits a practical left-to-right beam-search algorithm for carrying out inference. Experiments show that our model benefits from using cross-sentence context in the language model, and it outperforms existing document translation approaches.

Selecting Backtranslated Data from Multiple Sources for Improved Neural Machine Translation [Website][PDF]
Xabier Soto, Dimitar Shterionov, Alberto Poncelas, and Andy Way 13:00–14:00

Machine translation (MT) has benefited from using synthetic training data originating from translating monolingual corpora, a technique known as backtranslation. Combining backtranslated data from different sources has led to better results than when using such data in isolation. In this work we analyse the impact that data translated with rule-based, phrase-based statistical and neural MT systems has on new MT systems. We use a real-world low-resource use-case (Basque-to-Spanish in the clinical domain) as well as a high-resource language pair (German-to-English) to test different scenarios with backtranslation and employ data selection to optimise the synthetic corpora. We exploit different data selection strategies in order to reduce the amount of data used, while at the same time maintaining high-quality MT systems. We further tune the data selection method by taking into account the quality of the MT systems used for backtranslation and lexical diversity of the resulting corpora. Our experiments show that incorporating backtranslated data from different sources can be beneficial, and that availing of data selection can yield improved performance.

[CL] Unsupervised Word Translation with Adversarial Autoencoder [Website][PDF]
Tasnim Mohiuddin and Shafiq Joty 13:00–14:00

Crosslingual word embeddings learned from monolingual embeddings have a crucial role in many downstream tasks, ranging from machine translation to transfer learning. Adversarial training has shown impressive success in learning crosslingual embeddings and the associated word translation task without any parallel data by mapping monolingual embeddings to a shared space. However, recent work has shown superior performance for non-adversarial methods in more challenging language pairs. In this article, we investigate adversarial autoencoder for unsupervised word

translation and propose two novel extensions to it that yield more stable training and improved results. Our method includes regularization terms to enforce cycle consistency and input reconstruction, and puts the target encoders as an adversary against the corresponding discriminator. We use two types of refinement procedures sequentially after obtaining the trained encoders and mappings from the adversarial training, namely, refinement with Procrustes solution and refinement with symmetric re-weighting. Extensive experimentations with high- and low-resource languages from two different data sets show that our method achieves better performance than existing adversarial and non-adversarial approaches and is also competitive with the supervised system. Along with performing comprehensive ablation studies to understand the contribution of different components of our adversarial model, we also conduct a thorough analysis of the refinement procedures to understand their effects.

Session 8B: NLP Applications-7

Empowering Active Learning to Jointly Optimize System and User Demands [Website][PDF]
Ji-Ung Lee, Christian M. Meyer, and Iryna Gurevych 13:00–14:00

Existing approaches to active learning maximize the system performance by sampling unlabeled instances for annotation that yield the most efficient training. However, when active learning is integrated with an end-user application, this can lead to frustration for participating users, as they spend time labeling instances that they would not otherwise be interested in reading. In this paper, we propose a new active learning approach that jointly optimizes the seemingly counteracting objectives of the active learning system (training efficiently) and the user (receiving useful instances). We study our approach in an educational application, which particularly benefits from this technique as the system needs to rapidly learn to predict the appropriateness of an exercise to a particular user, while the users should receive only exercises that match their skills. We evaluate multiple learning strategies and user types with data from real users and find that our joint approach better satisfies both objectives when alternative methods lead to many unsuitable exercises for end users.

Encoder-Decoder Models Can Benefit from Pre-trained Masked Language Models in Grammatical Error Correction [Website][PDF]
Masahiro Kaneko, Masato Mita, Shun Kiyono, Jun Suzuki, and Kentaro Inui 13:00–14:00

This paper investigates how to effectively incorporate a pre-trained masked language model (MLM), such as BERT, into an encoder-decoder (EncDec) model for grammatical error correction (GEC). The answer to this question is not as straightforward as one might expect because the previous common methods for incorporating a MLM into an EncDec model have potential drawbacks when applied to GEC. For example, the distribution of the inputs to a GEC model can be considerably different (erroneous, clumsy, etc.) from that of the corpora used for pre-training MLMs; however, this issue is not addressed in the previous methods. Our experiments show that our proposed method, where we first fine-tune a MLM with a given GEC corpus and then use the output of the fine-tuned MLM as additional features in the GEC model, maximizes the benefit of the MLM. The best-performing model achieves state-of-the-art performances on the BEA-2019 and CoNLL-2014 benchmarks. Our code is publicly available at: https://github.com/kanekomasahiro/bert-gec.

Graph Neural News Recommendation with Unsupervised Preference Disentanglement [Website][PDF]
Linmei Hu, Siyong Xu, Chen Li, Cheng Yang, Chuan Shi, Nan Duan, Xing Xie, and Ming Zhou 13:00–14:00

With the explosion of news information, personalized news recommendation has become very important for users to quickly find their interested contents. Most existing methods usually learn the representations of users and news from news contents for recommendation. However, they seldom consider high-order connectivity underlying the user-news interactions. Moreover, existing methods failed to disentangle a user's latent preference factors which cause her clicks on different news. In this paper, we model the user-news interactions as a bipartite graph and propose a novel Graph Neural News Recommendation model with Unsupervised Preference Disentanglement, named GNUD. Our model can encode high-order relationships into user and news representations by information propagation along the graph. Furthermore, the learned representations are disentangled with latent preference factors by a neighborhood routing algorithm, which can enhance expressiveness and interpretability. A preference regularizer is also designed to force each disentangled subspace to independently reflect an isolated preference, improving the quality of the disentangled representations. Experimental results on real-world news datasets demonstrate that our proposed model can effectively improve the performance of news recommendation and outperform state-of-the-art news recommendation methods.

Hiring Now: A Skill-Aware Multi-Attention Model for Job Posting Generation [Website][PDF]
Liting Liu, Jie Liu, Wenzheng Zhang, Ziming Chi, Wenxuan Shi, and Yalou Huang 13:00–14:00

Writing a good job posting is a critical step in the recruiting process, but the task is often more difficult than many people think. It is challenging to specify the level of education, experience, relevant skills per the company information and job description. To this end, we propose a novel task of Job Posting Generation (JPG) which is cast as a conditional text generation problem to generate job requirements according to the job descriptions. To deal with this task, we devise a data-driven global Skill-Aware Multi-Attention generation model, named SAMA. Specifically, to model the complex mapping relationships between input and output, we design a hierarchical decoder that we first label the job description with multiple skills, then we generate a complete text guided by the skill labels. At the same time, to exploit the prior knowledge about the skills, we further construct a skill knowledge graph to capture the global prior knowledge of skills and refine the generated results. The proposed approach is evaluated on real-world job posting data. Experimental results clearly demonstrate the effectiveness of the proposed method.

Identifying Principals and Accessories in a Complex Case based on the Comprehension of Fact Description [Website][PDF]
Yakun Hu, Zhunchen Luo, and Wenhan Chao 13:00–14:00

In this paper, we study the problem of identifying the principals and accessories from the fact description with multiple defendants in a criminal case. We treat the fact descriptions as narrative texts and the defendants as roles over the narrative story. We propose to model the defendants with *behavioral semantic information* and *statistical characteristics*, then learning the importances of defendants within a learning-to-rank framework. Experimental results on a real-world dataset demonstrate the behavior analysis can effectively model the defendants' impacts in a complex case.

Joint Modelling of Emotion and Abusive Language Detection [Website][PDF]
Santhosh Rajamanickam, Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova 13:00–14:00

The rise of online communication platforms has been accompanied by some undesirable effects, such as the proliferation of aggressive and abusive behaviour online. Aiming to tackle this problem, the natural language processing (NLP) community has experimented with a range of techniques for abuse detection. While achieving substantial success, these methods have so far only focused on modelling the linguistic properties of the comments and the online communities of users, disregarding the emotional state of the users and how this might affect their language. The latter is, however, inextricably linked to abusive behaviour. In this paper, we present the first joint model of emotion and abusive language detection, experimenting in a multi-task learning framework that allows one task to inform the other. Our results demonstrate that incorporating affective features leads to significant improvements in abuse detection performance across datasets.

Programming in Natural Language with fuSE: Synthesizing Methods from Spoken Utterances Using Deep Natural Language Understanding [Website][PDF]

Sebastian Weigelt, Vanessa Steurer, Tobias Hey, and Walter F. Tichy 13:00–14:00

The key to effortless end-user programming is natural language. We examine how to teach intelligent systems new functions, expressed in natural language. As a first step, we collected 3168 samples of teaching efforts in plain English. Then we built fuSE, a novel system that translates English function descriptions into code. Our approach is three-tiered and each task is evaluated separately. We first classify whether an intent to teach new functionality is present in the utterance (accuracy: 97.7% using BERT). Then we analyze the linguistic structure and construct a semantic model (accuracy: 97.6% using a BiLSTM). Finally, we synthesize the signature of the method, map the intermediate steps (instructions in the method body) to API calls and inject control structures (F_1: 67.0% with information retrieval and knowledge-based methods). In an end-to-end evaluation on an unseen dataset fuSE synthesized 84.6% of the method signatures and 79.2% of the API calls correctly.

Toxicity Detection: Does Context Really Matter? [Website][PDF]

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos 13:00–14:00

Moderation is crucial to promoting healthy online discussions. Although several 'toxicity' detection datasets and models have been published, most of them ignore the context of the posts, implicitly assuming that comments may be judged independently. We investigate this assumption by focusing on two questions: (a) does context affect the human judgement, and (b) does conditioning on context improve performance of toxicity detection systems? We experiment with Wikipedia conversations, limiting the notion of context to the previous post in the thread and the discussion title. We find that context can both amplify or mitigate the perceived toxicity of posts. Moreover, a small but significant subset of manually labeled posts (5% in one of our experiments) end up having the opposite toxicity labels if the annotators are not provided with context. Surprisingly, we also find no evidence that context actually improves the performance of toxicity classifiers, having tried a range of classifiers and mechanisms to make them context aware. This points to the need for larger datasets of comments annotated in context. We make our code and data publicly available.

Session 8B: Resources and Evaluation-8

ASSET: A Dataset for Tuning and Evaluation of Sentence Simplification Models with Multiple Rewriting Transformations [Website][PDF]
Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia 13:00–14:00

In order to simplify a sentence, human editors perform multiple rewriting transformations: they split it into several shorter sentences, paraphrase words (i.e. replacing complex words or phrases by simpler synonyms), reorder components, and/or delete information deemed unnecessary. Despite these varied range of possible text alterations, current models for automatic sentence simplification are evaluated using datasets that are focused on a single transformation, such as lexical paraphrasing or splitting. This makes it impossible to understand the ability of simplification models in more realistic settings. To alleviate this limitation, this paper introduces ASSET, a new dataset for assessing sentence simplification in English. ASSET is a crowdsourced multi-reference corpus where each simplification was produced by executing several rewriting transformations. Through quantitative and qualitative experiments, we show that simplifications in ASSET are better at capturing characteristics of simplicity when compared to other standard evaluation datasets for the task. Furthermore, we motivate the need for developing better methods for automatic evaluation using ASSET, since we show that current popular metrics may not be suitable when multiple simplification transformations are performed.

Automatic Machine Translation Evaluation using Source Language Inputs and Cross-lingual Language Model [Website][PDF]
Kosuke Takahashi, Katsuhito Sudoh, and Satoshi Nakamura 13:00–14:00

We propose an automatic evaluation method of machine translation that uses source language sentences regarded as additional pseudo references. The proposed method evaluates a translation hypothesis in a regression model. The model takes the paired source, reference, and hypothesis sentence all together as an input. A pretrained large scale cross-lingual language model encodes the input to sentence-pair vectors, and the model predicts a human evaluation score with those vectors. Our experiments show that our proposed method using Cross-lingual Language Model (XLM) trained with a translation language modeling (TLM) objective achieves a higher correlation with human judgments than a baseline method that uses only hypothesis and reference sentences. Additionally, using source sentences in our proposed method is confirmed to improve the evaluation performance.

Fatality Killed the Cat or: BabelPic, a Multimodal Dataset for Non-Concrete Concepts [Website][PDF]
Agostina Calabrese, Michele Bevilacqua, and Roberto Navigli 13:00–14:00

Thanks to the wealth of high-quality annotated images available in popular repositories such as ImageNet, multimodal language-vision research is in full bloom. However, events, feelings and many other kinds of concepts which can be visually grounded are not well represented in current datasets. Nevertheless, we would expect a wide-coverage language understanding system to be able to classify images depicting recess and remorse, not just cats, dogs and bridges. We fill this gap by presenting BabelPic, a hand-labeled dataset built by cleaning the image-synset association found within the BabelNet Lexical Knowledge Base (LKB). BabelPic explicitly targets non-concrete concepts, thus providing refreshing new data for the community. We also show that pre-trained language-vision systems can be used to further expand the resource by exploiting natural language knowledge available in the LKB. BabelPic is available for download at http://babelpic.org.

[TACL] Paraphrase-Sense-Tagged Sentences [Website][PDF]
Anne Cocos and Chris Callison-Burch 13:00–14:00

Many natural language processing tasks require discriminating the particular meaning of a word in context, but building corpora for developing sense-aware models can be a challenge. We present a large resource of example usages for words having a particular meaning, called Paraphrase-Sense-Tagged Sentences (PSTS). Built upon the premise that a word's paraphrases instantiate its fine-grained meanings – i.e. 'bug' has different meanings corresponding to its paraphrases 'fly' and 'microbe' – the resource contains up to 10,000 sentences for each of 3 million target-paraphrase pairs where the target word takes on the meaning of the paraphrase. We describe an automatic method based on bilingual pivoting used to enumerate sentences for PSTS, and present two models for ranking PSTS sentences based on their quality. Finally, we demonstrate the utility of PSTS by using it to build a dataset for the task of hypernym prediction in context. Training a model on this automatically-generated dataset produces accuracy that is competitive with a model trained on smaller datasets crafted with some manual effort.

That is a Known Lie: Detecting Previously Fact-Checked Claims [Website][PDF]
Shaden Shaar, Nikolay Babulkov, Giovanni Da San Martino, and Preslav Nakov 13:00–14:00

The recent proliferation of "fake news" has triggered a number of responses, most notably the emergence of several manual fact-checking initiatives. As a result and over time, a large number of fact-checked claims have been accumulated, which increases the likelihood that a new claim in social media or a new statement by a politician might have already been fact-checked by some trusted fact-checking organization, as viral claims often come back after a while in social media, and politicians like to repeat their favorite statements, true or false, over and over again. As manual fact-checking is very time-consuming (and fully automatic fact-checking has credibility issues), it is important to try to save this effort and to avoid wasting time on claims that have already been fact-checked. Interestingly, despite the importance of the task, it has been largely ignored by the research community so far. Here, we aim to bridge this gap. In particular, we formulate the task and we discuss how it relates to, but also differs from, previous work. We further create a specialized dataset, which we release to the research community. Finally, we present learning-to-rank experiments that demonstrate sizable improvements over state-of-the-art retrieval and textual similarity approaches.

Session 8B: Sentiment Analysis, Stylistic Analysis, and Argument Mining-6

Aspect Sentiment Classification with Document-level Sentiment Preference Modeling [Website][PDF]
Xiao Chen, Changlong Sun, Jingjing Wang, Shoushan Li, Luo Si, Min Zhang, and Guodong Zhou 13:00–14:00

In the literature, existing studies always consider Aspect Sentiment Classification (ASC) as an independent sentence-level classification problem aspect by aspect, which largely ignore the document-level sentiment preference information, though obviously such information is crucial for alleviating the information deficiency problem in ASC. In this paper, we explore two kinds of sentiment preference information inside a document, i.e., contextual sentiment consistency w.r.t. the same aspect (namely intra-aspect sentiment consistency) and contextual sentiment tendency w.r.t. all the related aspects (namely inter-aspect sentiment tendency). On the basis, we propose a Cooperative Graph Attention Networks (CoGAN) approach for cooperatively learning the aspect-related sentence representation. Specifically, two graph attention networks are leveraged to model above two kinds of document-level sentiment preference information respectively, followed by an interactive mechanism to integrate the two-fold preference. Detailed evaluation demonstrates the great advantage of the proposed approach to ASC over the state-of-the-art baselines. This justifies the importance of the document-level sentiment preference information to ASC and the effectiveness of our approach capturing such information.

ECPE-2D: Emotion-Cause Pair Extraction based on Joint Two-Dimensional Representation, Interaction and Prediction [Website][PDF]
Zixiang Ding, Rui Xia, and Jianfei Yu 13:00–14:00

In recent years, a new interesting task, called emotion-cause pair extraction (ECPE), has emerged in the area of text emotion analysis. It aims at extracting the potential pairs of emotions and their corresponding causes in a document. To solve this task, the existing research employed a two-step framework, which first extracts individual emotion set and cause set, and then pair the corresponding emotions and causes. However, such a pipeline of two steps contains some inherent flaws: 1) the modeling does not aim at extracting the final emotion-cause pair directly; 2) the errors from the first step will affect the performance of the second step. To address these shortcomings, in this paper we propose a new end-to-end approach, called ECPE-Two-Dimensional (ECPE-2D), to represent the emotion-cause pairs by a 2D representation scheme. A 2D transformer module and two variants, window-constrained and cross-road 2D transformers, are further proposed to model the interactions of different emotion-cause pairs. The 2D representation, interaction, and prediction are integrated into a joint framework. In addition to the advantages of joint modeling, the experimental results on the benchmark emotion cause corpus show that our approach improves the F1 score of the state-of-the-art from 61.28% to 68.89%.

From Arguments to Key Points: Towards Automatic Argument Summarization [Website][PDF]
Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim 13:00–14:00

Generating a concise summary from a large collection of arguments on a given topic is an intriguing yet understudied problem. We propose to represent such summaries as a small set of talking points, termed *key points*, each scored according to its salience. We show, by analyzing a large dataset of crowd-contributed arguments, that a small number of key points per topic is typically sufficient for covering the vast majority of the arguments. Furthermore, we found that a domain expert can often predict these key points in advance. We study the task of argument-to-key point mapping, and introduce a novel large-scale dataset for this task. We report empirical results for an extensive set of experiments with this dataset, showing promising performance.

He said "who's gonna take care of your children when you are at ACL?": Reported Sexist Acts are Not Sexist [Website][PDF]
Patricia Chiril, Véronique MORICEAU, Farah Benamara, Alda Mari, Gloria Origgi, and Marlène Coulomb-Gully 13:00–14:00

In a context of offensive content mediation on social media now regulated by European laws, it is important not only to be able to automatically detect sexist content but also to identify if a message with a sexist content is really sexist or is a story of sexism experienced by a woman. We propose: (1) a new characterization of sexist content inspired by speech acts theory and discourse analysis studies, (2) the first French dataset annotated for sexism detection, and (3) a set of deep learning experiments trained on top of a combination of several tweet's vectorial representations (word embeddings, linguistic features, and various generalization strategies). Our results are encouraging and constitute a first step towards offensive content moderation.

Modeling Label Semantics for Predicting Emotional Reactions [Website][PDF]
Radhika Gaonkar, Heeyoung Kwon, Mohaddeseh Bastan, Niranjan Balasubramanian, and Nathanael Chambers 13:00–14:00

Predicting how events induce emotions in the characters of a story is typically seen as a standard multi-label classification task, which usually treats labels as anonymous classes to predict. They ignore information that may be conveyed by the emotion labels themselves. We propose that the semantics of emotion labels can guide a model's attention when representing the input story. Further, we observe that the emotions evoked by an event are often related: an event that evokes joy is unlikely to also evoke sadness. In this work, we explicitly model label classes via label embeddings, and add mechanisms that track label-label correlations both during training and inference. We also introduce a new semi-supervision strategy that regularizes for the correlations on unlabeled data. Our empirical evaluations show that modeling label semantics yields consistent benefits, and we advance the state-of-the-art on an emotion inference task.

SKEP: Sentiment Knowledge Enhanced Pre-training for Sentiment Analysis [Website] [PDF]
Hao Tian, Can Gao, Xinyan Xiao, Hao Liu, Bolei He, Hua Wu, Haifeng Wang, and feng wu feng 13:00–14:00

Recently, sentiment analysis has seen remarkable advance with the help of pre-training approaches. However, sentiment knowledge, such as sentiment words and aspect-sentiment pairs, is ignored in the process of pre-training, despite the fact that they are widely used in traditional sentiment analysis approaches. In this paper, we introduce Sentiment Knowledge Enhanced Pre-training (SKEP) in order to learn a unified sentiment representation for multiple sentiment analysis tasks. With the help of automatically-mined knowledge, SKEP conducts sentiment masking and constructs three sentiment knowledge prediction objectives, so as to embed sentiment information at the word, polarity and aspect level into pre-trained sentiment representation. In particular, the prediction of aspect-sentiment pairs is converted into multi-label classification, aiming to capture the dependency between words in a pair. Experiments on three kinds of sentiment tasks show that SKEP significantly outperforms strong pre-training baseline, and achieves new state-of-the-art results on most of the test datasets. We release our code at https://github.com/baidu/Senta.

Session 8B: Speech and Multimodality-5

How Accents Confound: Probing for Accent Information in End-to-End Speech Recognition Systems [Website][PDF]
Archiki Prasad and Preethi Jyothi 13:00–14:00

In this work, we present a detailed analysis of how accent information is reflected in the internal representation of speech in an end-to-end automatic speech recognition (ASR) system. We use a state-of-the-art end-to-end ASR system, comprising convolutional and recurrent layers, that is trained on a large amount of US-accented English speech and evaluate the model on speech samples from seven different English accents. We examine the effects of accent on the internal representation using three main probing techniques: a) Gradient-based explanation methods, b) Information-theoretic measures, and c) Outputs of accent and phone classifiers. We find different accents exhibiting similar trends irrespective of the probing technique used. We also find that most accent information is encoded within the first recurrent layer, which is suggestive of how one could adapt such an end-to-end model to learn representations that are invariant to accents.

Learning Spoken Language Representations with Neural Lattice Language Modeling [Website][PDF]
Chao-Wei Huang and Yun-Nung Chen 13:00–14:00

Pre-trained language models have achieved huge improvement on many NLP tasks. However, these methods are usually designed for written text, so they do not consider the properties of spoken language. Therefore, this paper aims at generalizing the idea of language model pre-training to lattices generated by recognition systems. We propose a framework that trains neural lattice language models to provide contextualized representations for spoken language understanding tasks. The proposed two-stage pre-training approach reduces the demands of speech data and has better efficiency. Experiments on intent detection and dialogue act recognition datasets demonstrate that our proposed method consistently outperforms strong baselines when evaluated on spoken inputs. The code is available at https://github.com/MiuLab/Lattice-ELMo.

Meta-Transfer Learning for Code-Switched Speech Recognition [Website][PDF]
Genta Indra Winata, Samuel Cahyawijaya, Zhaojiang Lin, Zihan Liu, Peng Xu, and Pascale Fung 13:00–14:00

An increasing number of people in the world today speak a mixed-language as a result of being multilingual. However, building a speech recognition system for code-switching remains difficult due to the availability of limited resources and the expense and significant effort required to collect mixed-language data. We therefore propose a new learning method, meta-transfer learning, to transfer learn on a code-switched speech recognition system in a low-resource setting by judiciously extracting information from high-resource monolingual datasets. Our model learns to recognize individual languages, and transfer them so as to better recognize mixed-language speech by conditioning the optimization on the code-switching data. Based on experimental results, our model outperforms existing baselines on speech recognition and language modeling tasks, and is faster to converge.

Multimodal Transformer for Multimodal Machine Translation [Website][PDF]
Shaowei Yao and Xiaojun Wan 13:00–14:00

Multimodal Machine Translation (MMT) aims to introduce information from other modality, generally static images, to improve the translation quality. Previous works propose various incorporation methods, but most of them do not consider the relative importance of multiple modalities. Equally treating all modalities may encode too much useless information from less important modalities. In this paper, we introduce the multimodal self-attention in Transformer to solve the issues above in MMT. The proposed method learns the representation of images based on the text, which avoids encoding irrelevant information in images. Experiments and visualization analysis demonstrate that our model benefits from visual information and substantially outperforms previous works and competitive baselines in terms of various metrics.

Sentiment and Emotion help Sarcasm? A Multi-task Learning Framework for Multi-Modal Sarcasm, Sentiment and Emotion Analysis [Website][PDF]
Dushyant Singh Chauhan, Dhanush S R, Asif Ekbal, and Pushpak Bhattacharyya 13:00–14:00

In this paper, we hypothesize that sarcasm is closely related to sentiment and emotion, and thereby propose a multi-task deep learning framework to solve all these three problems simultaneously in a multi-modal conversational scenario. We, at first, manually annotate the recently released multi-modal MUStARD sarcasm dataset with sentiment and emotion classes, both implicit and explicit. For multi-tasking, we propose two attention mechanisms, viz. Inter-segment Inter-modal Attention (Ie-Attention) and Intra-segment Inter-modal Attention (Ia-Attention). The main motivation of Ie-Attention is to learn the relationship between the different segments of the sentence across the modalities. In contrast, Ia-Attention focuses within the same segment of the sentence across the modalities. Finally, representations from both the attentions are concatenated and shared across the five classes (i.e., sarcasm, implicit sentiment, explicit sentiment, implicit emotion, explicit emotion) for multi-tasking. Experimental results on the extended version of the MUStARD dataset show the efficacy of our proposed approach for sarcasm detection over the existing state-of-the-art systems. The evaluation also shows that the proposed multi-task framework yields better performance for the primary task, i.e., sarcasm detection, with the help of two secondary tasks, emotion and sentiment analysis.

SimulSpeech: End-to-End Simultaneous Speech to Text Translation [Website][PDF]
Yi Ren, Jinglin Liu, Xu Tan, Chen Zhang, Tao QIN, Zhou Zhao, and Tie-Yan Liu 13:00–14:00

In this work, we develop SimulSpeech, an end-to-end simultaneous speech to text translation system which translates speech in source language to text in target language concurrently. SimulSpeech consists of a speech encoder, a speech segmenter and a text decoder, where 1) the segmenter builds upon the encoder and leverages a connectionist temporal classification (CTC) loss to split the input streaming speech in real time, 2) the encoder-decoder attention adopts a wait-k strategy for simultaneous translation. SimulSpeech is more challenging than previous cascaded systems (with simultaneous automatic speech recognition (ASR) and simultaneous neural machine translation (NMT)). We introduce two novel knowledge distillation methods to ensure the performance: 1) Attention-level knowledge distillation transfers the knowledge from the multiplication of the attention matrices of simultaneous NMT and ASR models to help the training of the attention mechanism in SimulSpeech; 2) Data-level knowledge distillation transfers the knowledge from the full-sentence NMT model and also reduces the complexity of data distribution to help on the optimization of SimulSpeech. Experiments on MuST-C English-Spanish and English-German spoken language translation datasets show that SimulSpeech achieves reasonable BLEU scores and lower delay compared to full-sentence end-to-end speech to text translation (without simultaneous translation), and better performance than the two-stage cascaded simultaneous translation model in terms of BLEU scores and translation delay.

Towards Emotion-aided Multi-modal Dialogue Act Classification [Website][PDF]
Tulika Saha, Aditya Patra, Sriparna Saha, and Pushpak Bhattacharyya 13:00–14:00

The task of Dialogue Act Classification (DAC) that purports to capture communicative intent has been studied extensively. But these studies limit themselves to text. Non-verbal features (change of tone, facial expressions etc.) can provide cues to identify DAs, thus stressing the benefit of incorporating multi-modal inputs in the task. Also, the emotional state of the speaker has a substantial effect on the choice of the dialogue act, since conversations are often influenced by emotions. Hence, the effect of emotion too on automatic identification of DAs needs to be studied. In this work, we address the role of *both* multi-modality and emotion recognition (ER) in DAC. DAC and ER help each other by way of multi-task learning. One of the major contributions of this work is a new dataset- multimodal Emotion aware Dialogue Act dataset called EMOTyDA, collected from open-sourced dialogue datasets. To demonstrate the utility of EMOTyDA, we build an attention based (self, inter-modal, inter-task) multi-modal, multi-task Deep Neural Network (DNN) for joint learning of DAs and emotions. We show empirically that multi-modality and multi-tasking achieve better performance of DAC compared to uni-modal and single task DAC variants.

Session 8B: Student Research Workshop

Pre-training via Leveraging Assisting Languages for Neural Machine Translation [Website][PDF]
Haiyue Song, Raj Dabre, Zhuoyuan Mao, Fei Cheng, Sadao Kurohashi, and Eiichiro Sumita 13:00–14:00

Sequence-to-sequence (S2S) pre-training using large monolingual data is known to improve performance for various S2S NLP tasks. However, large monolingual corpora might not always be available for the languages of interest (LOI). Thus, we propose to exploit monolingual corpora of other languages to complement the scarcity of monolingual corpora for the LOI. We utilize script mapping (Chinese to Japanese) to increase the similarity (number of cognates) between the monolingual corpora of helping languages and LOI. An empirical case study of low-resource Japanese-English neural machine translation (NMT) reveals that leveraging large Chinese and French monolingual corpora can help overcome the shortage of Japanese and English monolingual corpora, respectively, for S2S pre-training. Using only Chinese and French monolingual corpora, we were able to improve Japanese-English translation quality by up to 8.5 BLEU in low-resource scenarios.

Preventing Critical Scoring Errors in Short Answer Scoring with Confidence Estimation [Website][PDF]
Hiroaki Funayama, Shota Sasaki, Yuichiroh Matsubayashi, Tomoya Mizumoto, Jun Suzuki, Masato Mita, and Kentaro Inui 13:00–14:00

Many recent Short Answer Scoring (SAS) systems have employed Quadratic Weighted Kappa (QWK) as the evaluation measure of their systems. However, we hypothesize that QWK is unsatisfactory for the evaluation of the SAS systems when we consider measuring their effectiveness in actual usage. We introduce a new task formulation of SAS that matches the actual usage. In our formulation, the SAS systems should extract as many scoring predictions that are not critical scoring errors (CSEs). We conduct the experiments in our new task formulation and demonstrate that a typical SAS system can predict scores with zero CSE for approximately 50% of test data at maximum by filtering out low-reliablility predictions on the basis of a certain confidence estimation. This result directly indicates the possibility of reducing half the scoring cost of human raters, which is more preferable for the evaluation of SAS systems.

Demo Session 3C

Time: *13:30–14:15*

exBERT: A Visual Analysis Tool to Explore Learned Representations in Transformer Models [Website][PDF]

Benjamin Hoover, Hendrik Strobelt, and Sebastian Gehrmann

Large Transformer-based language models can route and reshape complex information via their multi-headed attention mechanism. Although the attention never receives explicit supervision, it can exhibit recognizable patterns following linguistic or positional information. Analyzing the learned representations and attentions is paramount to furthering our understanding of the inner workings of these models. However, analyses have to catch up with the rapid release of new models and the growing diversity of investigation techniques. To support analysis for a wide variety of models, we introduce exBERT, a tool to help humans conduct flexible, interactive investigations and formulate hypotheses for the model-internal reasoning process. exBERT provides insights into the meaning of the contextual representations and attention by matching a human-specified input to similar contexts in large annotated datasets. By aggregating the annotations of the matched contexts, exBERT can quickly replicate findings from literature and extend them to previously not analyzed models.

Demo Session 4A

Time: *17:00–17:45*

Nakdan: Professional Hebrew Diacritizer [Website][PDF]
Avi Shmidman, Shaltiel Shmidman, Moshe Koppel, and Yoav Goldberg

We present a system for automatic diacritization of Hebrew Text. The system combines modern neural models with carefully curated declarative linguistic knowledge and comprehensive manually constructed tables and dictionaries. Besides providing state of the art diacritization accuracy, the system also supports an interface for manual editing and correction of the automatic output, and has several features which make it particularly useful for preparation of scientific editions of historical Hebrew texts. The system supports Modern Hebrew, Rabbinic Hebrew and Poetic Hebrew. The system is freely accessible for all use at http://nakdanpro.dicta.org.il

Photon: A Robust Cross-Domain Text-to-SQL System [Website][PDF]
Jichuan Zeng, Xi Victoria Lin, Steven C.H. Hoi, Richard Socher, Caiming Xiong, Michael Lyu, and Irwin King

Natural language interfaces to databases(NLIDB) democratize end user access to relational data. Due to fundamental differences between natural language communication and programming, it is common for end users to issue questions that are ambiguous to the system or fall outside the semantic scope of its underlying query language. We present PHOTON, a robust, modular, cross-domain NLIDB that can flag natural language input to which a SQL mapping cannot be immediately determined. PHOTON consists of a strong neural semantic parser (63.2% structure accuracy on the Spider dev benchmark), a human-in-the-loop question corrector, a SQL executor and a response generator. The question corrector isa discriminative neural sequence editor which detects confusion span(s) in the input question and suggests rephrasing until a translatable input is given by the user or a maximum number of iterations are conducted. Experiments on simulated data show that the proposed method effectively improves the robustness of text-to-SQL system against untranslatable user input.The live demo of our system is available at http://www.naturalsql.com

OpusFilter: A Configurable Parallel Corpus Filtering Toolbox [Website][PDF]
Mikko Aulamo, Sami Virpioja, and Jörg Tiedemann

This paper introduces OpusFilter, a flexible and modular toolbox for filtering parallel corpora. It implements a number of components based on heuristic filters, language identification libraries, character-based language models, and word alignment tools, and it can easily be extended with custom filters. Bitext segments can be ranked according to their quality or domain match using single features or a logistic regression model that can be trained without manually labeled training data. We demonstrate the effectiveness of OpusFilter on the example of a Finnish-English news translation task based on noisy web-crawled training data. Applying our tool leads to improved translation quality while significantly reducing the size of the training data, also clearly outperforming an alternative ranking given in the crawled data set. Furthermore, we show the ability of OpusFilter to perform data selection for domain adaptation.

Session 9A Overview – Tuesday, July 7, 2020 17:00–18:00

Track A *Dialogue and Interactive Systems-11* Abstracts	CraftAssist Instruction Parsing: Semantic Parsing for a Voxel-World Assistant *Srinet, Jernite, Gray, and* [Website][PDF]	Don't Say That! Making Inconsistent Dialogue Unlikely with Unlikelihood Training *Li, Roller, Kulikov, Welleck, Boureau, Cho, and Weston* [Website][PDF]			
Track B *Interpretability and Analysis of Models for NLP-5* Abstracts	Compositionality and Generalization In Emergent Languages *Chaabouni, Kharitonov, Bouchacourt, Dupoux, and Baroni* [Website][PDF]	How does BERT's attention change when you fine-tune? An analysis methodology and a case study in negation scope *Zhao and Bethard* [Website][PDF]	Human Attention Maps for Text Classification: Do Humans and Neural Networks Focus on the Same Words? *Sen, Hartvigsen, Yin, Kong, and Rundensteiner* [Website][PDF]	Influence Paths for Characterizing Subject-Verb Number Agreement in LSTM Language Models *Lu, Mardziel, Leino, Fredrikson, and Datta* [Website][PDF]	Information-Theoretic Probing for Linguistic Structure *Pimentel, Valvoda, Hall Maudslay, Zmigrod, Williams, and Cotterell* [Website][PDF]
	Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings *Bommasani, Davis, and Cardie* [Website][PDF]	Learning to Deceive with Attention-Based Explanations *Pruthi, Gupta, Dhingra, Neubig, and Lipton* [Website][PDF]	On the Spontaneous Emergence of Discrete and Compositional Signals *Geffen Lan, Chemla, and Steinert-Threlkeld* [Website][PDF]	Similarity Analysis of Contextual Word Representation Models *Wu, Belinkov, Sajjad, Durrani, Dalvi, and Glass* [Website][PDF]	Spying on Your Neighbors: Fine-grained Probing of Contextual Embeddings for Information about Surrounding Words *Klafka and Ettinger* [Website][PDF]
Track C *Language Grounding to Vision, Robotics and Beyond-3* Abstracts	Dense-Caption Matching and Frame-Selection Gating for Temporal Localization in VideoQA *Kim, Tang, and Bansal* [Website][PDF]	Shaping Visual Representations with Language for Few-Shot Classification *Mu, Liang, and Goodman* [Website][PDF]			
Track D *Machine Learning for NLP-10* Abstracts	A Probabilistic Generative Model for Typographical Analysis of Early Modern Printing *Goyal, Dyer, Warren, G'Sell, and Berg-Kirkpatrick* [Website][PDF]	Discrete Latent Variable Representations for Low-Resource Text Classification *Jin, Wiseman, Stratos, and Livescu* [Website][PDF]	Learning Constraints for Structured Prediction Using Rectifier Networks *Pan, Mehta, and Srikumar* [Website][PDF]	Pretraining with Contrastive Sentence Objectives Improves Discourse Performance of Language Models *Iter, Guu, Lansing, and Jurafsky* [Website][PDF]	SenseBERT: Driving Some Sense into BERT *Levine, Lenz, Dagan, Ram, Padnos, Sharir, Shalev-Shwartz, Shashua, and Shoham* [Website][PDF]
	[TACL] SpanBERT: Improving Pre-training by Representing and Predicting Spans *Joshi, Chen, Liu, Weld, Zettlemoyer, and Levy* [Website][PDF]				

Track E *Resources and Evaluation-9* Abstracts	A Recipe for Creating Multimodal Aligned Datasets for Sequential Tasks *Lin, Rao, Celikyilmaz, Nouri, Brockett, Dey, and Dolan* [Website][PDF]	ASSET: A Dataset for Tuning and Evaluation of Sentence Simplification Models with Multiple Rewriting Transformations *Alva-Manchego, Martin, Bordes, Scarton, Sagot, and Specia* [Website][PDF]	Adversarial NLI: A New Benchmark for Natural Language Understanding *Nie, Williams, Dinan, Bansal, Weston, and Kiela* [Website][PDF]	Beyond Accuracy: Behavioral Testing of NLP Models with CheckList *Ribeiro, Wu, Guestrin, and Singh* [Website][PDF]	ChartDialogs: Plotting from Natural Language Instructions *Shao and Nakashole* [Website][PDF]
	Code and Named Entity Recognition in StackOverflow *Tabassum, Maddela, Xu, and Ritter* [Website][PDF]	Dialogue-Based Relation Extraction *Yu, Sun, Cardie, and Yu* [Website][PDF]	Facet-Aware Evaluation for Extractive Summarization *Mao, Liu, Zhu, Ren, and Han* [Website][PDF]	Fatality Killed the Cat or: BabelPic, a Multimodal Dataset for Non-Concrete Concepts *Calabrese, Bevilacqua, and Navigli* [Website][PDF]	[CL] LINSPECTOR: Multilingual Probing Tasks for Word Representations *Şahin, Vania, Kuznetsov, and Gurevych* [Website][PDF]
	More Diverse Dialogue Datasets via Diversity-Informed Data Collection *Stasaski, Yang, and Hearst* [Website][PDF]	ParaCrawl: Web-Scale Acquisition of Parallel Corpora *Bañón, Chen, Haddow, Heafield, Hoang, Esplà-Gomis, Forcada, Kamran, Kirefu, Koehn, Ortiz Rojas, Pla Sempere, Ramírez-Sánchez, Sarrías, Strelec, Thompson, Waites, Wiggins, and Zaragoza* [Website][PDF]	S2ORC: The Semantic Scholar Open Research Corpus *Lo, Wang, Neumann, Kinney, and Weld* [Website][PDF]	Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics *Mathur, Baldwin, and Cohn* [Website][PDF]	
Track F *Student Research Workshop* Abstracts	Checkpoint Reranking: An Approach to Select Better Hypothesis for Neural Machine Translation Systems *Pandramish and Sharma* [Website][PDF]	Cross-Lingual Disaster-related Multi-label Tweet Classification with Manifold Mixup *Ray Chowdhury, Caragea, and Caragea* [Website][PDF]	Inducing Grammar from Long Short-Term Memory Networks by Shapley Decomposition *Zhang and Nie* [Website][PDF]	Exploring the Role of Context to Distinguish Rhetorical and Information-Seeking Questions *Zhuang and Riloff* [Website][PDF]	
Track G *Summarization-4* Abstracts	A Transformer-based Approach for Source Code Summarization *Ahmad, Chakraborty, Ray, and Chang* [Website][PDF]	Asking and Answering Questions to Evaluate the Factual Consistency of Summaries *Wang, Cho, and Lewis* [Website][PDF]	Discourse-Aware Neural Extractive Text Summarization *Xu, Gan, Cheng, and Liu* [Website][PDF]	Discrete Optimization for Unsupervised Sentence Summarization with Word-Level Extraction *Schumann, Mou, Lu, Vechtomova, and Markert* [Website][PDF]	Exploring Content Selection in Summarization of Novel Chapters *Ladhak, Li, Al-Onaizan, and McKeown* [Website][PDF]

	FEQA: A Question Answering Evaluation Framework for Faithfulness Assessment in Abstractive Summarization *Durmus, He, and Diab* [Website][PDF]	Fact-based Content Weighting for Evaluating Abstractive Summarisation *Xu, Dušek, Li, Rieser, and Konstas* [Website][PDF]	Hooks in the Headline: Learning to Generate Headlines with Controlled Styles *Jin, Jin, Zhou, Orii, and Szolovits* [Website][PDF]	Knowledge Graph-Augmented Abstractive Summarization with Semantic-Driven Cloze Reward *Huang, Wu, and Wang* [Website][PDF]	Optimizing the Factual Correctness of a Summary: A Study of Summarizing Radiology Reports *Zhang, Merck, Tsai, Manning, and Langlotz* [Website][PDF]
	Storytelling with Dialogue: A Critical Role Dungeons and Dragons Dataset *Rameshkumar and Bailey* [Website][PDF]	The Summary Loop: Learning to Write Abstractive Summaries Without Examples *Laban, Hsi, Canny, and Hearst* [Website][PDF]	Unsupervised Opinion Summarization as Copycat-Review Generation *Bražinskas, Lapata, and Titov* [Website][PDF]		
Track H *Theme-1* Abstracts	(Re)construing Meaning in NLP *Trott, Timponi Torrent, Chang, and Schneider* [Website][PDF]	Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data *Bender and Koller* [Website][PDF]	Examining Citations of Natural Language Processing Literature *Mohammad* [Website][PDF]	How Can We Accelerate Progress Towards Human-like Linguistic Generalization? *Linzen* [Website][PDF]	How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence *Zhong, Xiao, Tu, Zhang, Liu, and Sun* [Website][PDF]
	Intermediate-Task Transfer Learning with Pretrained Language Models: When and Why Does It Work? *Pruksachatkun, Phang, Liu, Htut, Zhang, Pang, Vania, Kann, and Bowman* [Website][PDF]	Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview *Shah, Schwartz, and Hovy* [Website][PDF]	What Does BERT with Vision Look At? *Li, Yatskar, Yin, Hsieh, and Chang* [Website][PDF]		

Session 9A Details

Session 9A: Dialogue and Interactive Systems-11

CraftAssist Instruction Parsing: Semantic Parsing for a Voxel-World Assistant [Website][PDF]
Kavya Srinet, Yacine Jernite, Jonathan Gray, and arthur szlam arthur 17:00–18:00

We propose a semantic parsing dataset focused on instruction-driven communication with an agent in the game Minecraft. The dataset consists of 7K human utterances and their corresponding parses. Given proper world state, the parses can be interpreted and executed in game. We report the performance of baseline models, and analyze their successes and failures.

Don't Say That! Making Inconsistent Dialogue Unlikely with Unlikelihood Training [Website][PDF]
Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston 17:00–18:00

Generative dialogue models currently suffer from a number of problems which standard maximum likelihood training does not address. They tend to produce generations that (i) rely too much on copying from the context, (ii) contain repetitions within utterances, (iii) overuse frequent words, and (iv) at a deeper level, contain logical flaws.In this work we show how all of these problems can be addressed by extending the recently introduced unlikelihood loss (Welleck et al., 2019) to these cases. We show that appropriate loss functions which regularize generated outputs to match human distributions are effective for the first three issues. For the last important general issue, we show applying unlikelihood to collected data of what a model should not do is effective for improving logical consistency, potentially paving the way to generative models with greater reasoning ability. We demonstrate the efficacy of our approach across several dialogue tasks.

Session 9A: Interpretability and Analysis of Models for NLP-5

Compositionality and Generalization In Emergent Languages [Website][PDF]
Rahma Chaabouni, Eugene Kharitonov, Diane Bouchacourt, Emmanuel Dupoux, and Marco Baroni
17:00–18:00

Natural language allows us to refer to novel composite concepts by combining expressions denoting their parts according to systematic rules, a property known as compositionality. In this paper, we study whether the language emerging in deep multi-agent simulations possesses a similar ability to refer to novel primitive combinations, and whether it accomplishes this feat by strategies akin to human-language compositionality. Equipped with new ways to measure compositionality in emergent languages inspired by disentanglement in representation learning, we establish three main results: First, given sufficiently large input spaces, the emergent language will naturally develop the ability to refer to novel composite concepts. Second, there is no correlation between the degree of compositionality of an emergent language and its ability to generalize. Third, while compositionality is not necessary for generalization, it provides an advantage in terms of language transmission: The more compositional a language is, the more easily it will be picked up by new learners, even when the latter differ in architecture from the original agents. We conclude that compositionality does not arise from simple generalization pressure, but if an emergent language does chance upon it, it will be more likely to survive and thrive.

How does BERT's attention change when you fine-tune? An analysis methodology and a case study in negation scope [Website][PDF]
Yiyun Zhao and Steven Bethard 17:00–18:00

Large pretrained language models like BERT, after fine-tuning to a downstream task, have achieved high performance on a variety of NLP problems. Yet explaining their decisions is difficult despite recent work probing their internal representations. We propose a procedure and analysis methods that take a hypothesis of how a transformer-based model might encode a linguistic phenomenon, and test the validity of that hypothesis based on a comparison between knowledge-related downstream tasks with downstream control tasks, and measurement of cross-dataset consistency. We apply this methodology to test BERT and RoBERTa on a hypothesis that some attention heads will consistently attend from a word in negation scope to the negation cue. We find that after fine-tuning BERT and RoBERTa on a negation scope task, the average attention head improves its sensitivity to negation and its attention consistency across negation datasets compared to the pre-trained models. However, only the base models (not the large models) improve compared to a control task, indicating there is evidence for a shallow encoding of negation only in the base models.

Human Attention Maps for Text Classification: Do Humans and Neural Networks Focus on the Same Words? [Website][PDF]
Cansu Sen, Thomas Hartvigsen, Biao Yin, Xiangnan Kong, and Elke Rundensteiner 17:00–18:00

Motivated by human attention, computational attention mechanisms have been designed to help neural networks adjust their focus on specific parts of the input data. While attention mechanisms are claimed to achieve interpretability, little is known about the actual relationships between machine and human attention. In this work, we conduct the first quantitative assessment of human versus computational attention mechanisms for the text classification task. To achieve this, we design and conduct a large-scale crowd-sourcing study to collect human attention maps that encode the parts of a text that humans focus on when conducting text classification. Based on this new resource of human attention dataset for text classification, YELP-HAT, collected on the publicly available YELP dataset, we perform a quantitative comparative analysis of machine attention maps created by deep learning models and human attention maps. Our analysis offers insights into the relationships between human versus machine attention maps along three dimensions: overlap in word selections, distribution over lexical categories, and context-dependency of sentiment polarity. Our findings open promising future research opportunities ranging from supervised attention to the design of human-centric attention-based explanations.

Influence Paths for Characterizing Subject-Verb Number Agreement in LSTM Language Models [Website][PDF]
Kaiji Lu, Piotr Mardziel, Klas Leino, Matt Fredrikson, and Anupam Datta 17:00–18:00

LSTM-based recurrent neural networks are the state-of-the-art for many natural language processing (NLP) tasks. Despite their performance, it is unclear whether, or how, LSTMs learn structural features of natural languages such as subject-verb number agreement in English. Lacking this understanding, the generality of LSTM performance on this task and their suitability for related tasks remains uncertain. Further, errors cannot be properly attributed to a lack of structural capability, training data omissions, or other exceptional faults. We introduce *influence paths*, a causal account of structural properties as carried by paths across gates and neurons of a recurrent neural network. The approach refines the notion of influence (the subject's grammatical number has influence on the grammatical number of the subsequent verb) into a set of gate or neuron-level paths. The set localizes and segments the concept (e.g., subject-verb agreement), its constituent elements (e.g., the subject), and related or interfering elements (e.g., attractors). We exemplify the methodology on a widely-studied multi-layer LSTM language model, demonstrating its accounting for subject-verb number agreement. The results offer both a finer and a more complete view of an LSTM's handling of this structural aspect of the English language than prior results based on diagnostic classifiers and ablation.

Information-Theoretic Probing for Linguistic Structure [Website][PDF]
Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell
17:00–18:00

The success of neural networks on a diverse set of NLP tasks has led researchers to question how much these networks actually "know" about natural language. Probes are a natural way of assessing this. When probing, a researcher chooses a linguistic task and trains a supervised model to predict annotations in that linguistic task from the network's learned representations. If the probe does well, the researcher may conclude that the representations encode knowledge related to the task. A commonly held belief is that using simpler models as probes is better; the logic is that simpler models will identify linguistic structure, but not learn the task itself. We propose an information-theoretic operationalization of probing as estimating mutual information that contradicts this received wisdom: one should always select the highest performing probe one can, even if it is more complex, since it will result in a tighter estimate, and thus reveal more of the linguistic information inherent in the representation. The experimental portion of our paper focuses on empirically estimating the mutual information between a linguistic property and BERT, comparing these estimates to several baselines. We evaluate on a set of ten typologically diverse languages often underrepresented in NLP research—plus English—totalling eleven languages. Our implementation is available in https://github.com/rycolab/info-theoretic-probing.

Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings [Website][PDF]
Rishi Bommasani, Kelly Davis, and Claire Cardie 17:00–18:00

Contextualized representations (e.g. ELMo, BERT) have become the default pretrained representations for downstream NLP applications. In some settings, this transition has rendered their static embedding predecessors (e.g. Word2Vec, GloVe) obsolete. As a side-effect, we observe that older interpretability methods for static embeddings — while more diverse and mature than those available for their dynamic counterparts — are underutilized in studying newer contextualized representations. Consequently, we introduce simple and fully general methods for converting from contextualized representations to static lookup-table embeddings which we apply to 5 popular pretrained models and 9 sets of pretrained weights. Our analysis of the resulting static embeddings notably reveals that pooling over many contexts significantly improves representational quality under intrinsic evaluation. Complementary to analyzing representational quality, we consider social biases encoded in pretrained representations with respect to gender, race/ethnicity, and religion and find that bias is encoded disparately across pretrained models and internal layers even for models with the same training data. Concerningly, we find dramatic inconsistencies between social bias estimators for word embeddings.

Learning to Deceive with Attention-Based Explanations [Website][PDF]
Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C. Lipton 17:00–18:00

Attention mechanisms are ubiquitous components in neural architectures applied to natural language processing. In addition to yielding gains in predictive accuracy, attention weights are often claimed to confer interpretability, purportedly useful both for providing insights to practitioners and for explaining why a model makes its decisions to stakeholders. We call the latter use of attention mechanisms into question by demonstrating a simple method for training models to produce deceptive attention masks. Our method diminishes the total weight assigned to designated impermissible tokens, even when the models can be shown to nevertheless rely on these features to drive predictions. Across multiple models and tasks, our approach manipulates attention weights while paying surprisingly little cost in accuracy. Through a human study, we show that our manipulated attention-based explanations deceive people into thinking that predictions from a model biased against gender minorities do not rely on the gender. Consequently, our results cast doubt on attention's reliability as a tool for auditing algorithms in the context of fairness and accountability.

On the Spontaneous Emergence of Discrete and Compositional Signals [Website][PDF]
Nur Geffen Lan, Emmanuel Chemla, and Shane Steinert-Threlkeld 17:00–18:00

We propose a general framework to study language emergence through signaling games with neural agents. Using a continuous latent space, we are able to (i) train using backpropagation, (ii) show that discrete messages nonetheless naturally emerge. We explore whether categorical perception effects follow and show that the messages are not compositional.

Similarity Analysis of Contextual Word Representation Models [Website][PDF]
John Wu, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass 17:00–18:00

This paper investigates contextual word representation models from the lens of similarity analysis. Given a collection of trained models, we measure the similarity of their internal representations and attention. Critically, these models come from vastly different architectures. We use existing and novel similarity measures that aim to gauge the level of localization of information in the deep models, and facilitate the investigation of which design factors affect model similarity, without requiring any external linguistic annotation. The analysis reveals that models within the same family are more similar to one another, as may be expected. Surprisingly, different architectures have rather similar representations, but different individual neurons. We also observed differences in information localization in lower and higher layers and found that higher layers are more affected by fine-tuning on downstream tasks.

Spying on Your Neighbors: Fine-grained Probing of Contextual Embeddings for Information about Surrounding Words [Website][PDF]
Josef Klafka and Allyson Ettinger 17:00–18:00

Although models using contextual word embeddings have achieved state-of-the-art results on a host of NLP tasks, little is known about exactly what information these embeddings encode about the context words that they are understood to reflect. To address this question, we introduce a suite of probing tasks that enable fine-grained testing of contextual embeddings for encoding of information about surrounding words. We apply these tasks to examine the popular BERT, ELMo and GPT contextual encoders, and find that each of our tested information types is indeed

encoded as contextual information across tokens, often with near-perfect recoverability—but the encoders vary in which features they distribute to which tokens, how nuanced their distributions are, and how robust the encoding of each feature is to distance. We discuss implications of these results for how different types of models break down and prioritize word-level context information when constructing token embeddings.

Session 9A: Language Grounding to Vision, Robotics and Beyond-3

Dense-Caption Matching and Frame-Selection Gating for Temporal Localization in VideoQA [Website][PDF]
Hyounghun Kim, Zineng Tang, and Mohit Bansal 17:00–18:00

Videos convey rich information. Dynamic spatio-temporal relationships between people/objects, and diverse multimodal events are present in a video clip. Hence, it is important to develop automated models that can accurately extract such information from videos. Answering questions on videos is one of the tasks which can evaluate such AI abilities. In this paper, we propose a video question answering model which effectively integrates multi-modal input sources and finds the temporally relevant information to answer questions. Specifically, we first employ dense image captions to help identify objects and their detailed salient regions and actions, and hence give the model useful extra information (in explicit textual format to allow easier matching) for answering questions. Moreover, our model is also comprised of dual-level attention (word/object and frame level), multi-head self/cross-integration for different sources (video and dense captions), and gates which pass more relevant information to the classifier. Finally, we also cast the frame selection problem as a multi-label classification task and introduce two loss functions, In-andOut Frame Score Margin (IOFSM) and Balanced Binary Cross-Entropy (BBCE), to better supervise the model with human importance annotations. We evaluate our model on the challenging TVQA dataset, where each of our model components provides significant gains, and our overall model outperforms the state-of-the-art by a large margin (74.09% versus 70.52%). We also present several word, object, and frame level visualization studies.

Shaping Visual Representations with Language for Few-Shot Classification [Website][PDF]
Jesse Mu, Percy Liang, and Noah Goodman 17:00–18:00

By describing the features and abstractions of our world, language is a crucial tool for human learning and a promising source of supervision for machine learning models. We use language to improve few-shot visual classification in the underexplored scenario where natural language task descriptions are available during training, but unavailable for novel tasks at test time. Existing models for this setting sample new descriptions at test time and use those to classify images. Instead, we propose language-shaped learning (LSL), an end-to-end model that regularizes visual representations to predict language. LSL is conceptually simpler, more data efficient, and outperforms baselines in two challenging few-shot domains.

Session 9A: Machine Learning for NLP-10

A Probabilistic Generative Model for Typographical Analysis of Early Modern Printing [Website] [PDF]
Kartik Goyal, Chris Dyer, Christopher Warren, Maxwell G'Sell, and Taylor Berg-Kirkpatrick 17:00–18:00

We propose a deep and interpretable probabilistic generative model to analyze glyph shapes in printed Early Modern documents. We focus on clustering extracted glyph images into underlying templates in the presence of multiple confounding sources of variance. Our approach introduces a neural editor model that first generates well-understood printing phenomena like spatial perturbations from template parameters via interpertable latent variables, and then modifies the result by generating a non-interpretable latent vector responsible for inking variations, jitter, noise from the archiving process, and other unforeseen phenomena associated with Early Modern printing. Critically, by introducing an inference network whose input is restricted to the visual residual between the observation and the interpretably-modified template, we are able to control and isolate what the vector-valued latent variable captures. We show that our approach outperforms rigid interpretable clustering baselines (c.f. Ocular) and overly-flexible deep generative models (VAE) alike on the task of completely unsupervised discovery of typefaces in mixed-fonts documents.

Discrete Latent Variable Representations for Low-Resource Text Classification [Website] [PDF]
Shuning Jin, Sam Wiseman, Karl Stratos, and Karen Livescu 17:00–18:00

While much work on deep latent variable models of text uses continuous latent variables, discrete latent variables are interesting because they are more interpretable and typically more space efficient. We consider several approaches to learning discrete latent variable models for text in the case where exact marginalization over these variables is intractable. We compare the performance of the learned representations as features for low-resource document and sentence classification. Our best models outperform the previous best reported results with continuous representations in these low-resource settings, while learning significantly more compressed representations. Interestingly, we find that an amortized variant of Hard EM performs particularly well in the lowest-resource regimes.

Learning Constraints for Structured Prediction Using Rectifier Networks [Website] [PDF]
Xingyuan Pan, Maitrey Mehta, and Vivek Srikumar 17:00–18:00

Various natural language processing tasks are structured prediction problems where outputs are constructed with multiple interdependent decisions. Past work has shown that domain knowledge, framed as constraints over the output space, can help improve predictive accuracy. However, designing good constraints often relies on domain expertise. In this paper, we study the problem of learning such constraints. We frame the problem as that of training a two-layer rectifier network to identify valid structures or substructures, and show a construction for converting a trained network into a system of linear constraints over the inference variables. Our experiments on several NLP tasks show that the learned constraints can improve the prediction accuracy, especially when the number of training examples is small.

Pretraining with Contrastive Sentence Objectives Improves Discourse Performance of Language Models [Website] [PDF]
Dan Iter, Kelvin Guu, Larry Lansing, and Dan Jurafsky 17:00–18:00

Recent models for unsupervised representation learning of text have employed a number of techniques to improve contextual word representations but have put little focus on discourse-level representations. We propose Conpono, an inter-sentence objective for pretraining language models that models discourse coherence and the distance between sentences. Given an anchor sentence, our model is trained to predict the text k sentences away using a sampled-softmax objective where the candidates consist of neighboring sentences and sentences randomly sampled from the corpus. On the discourse representation benchmark DiscoEval, our model improves over the previous state-of-the-art by up to 13% and on average 4% absolute across 7 tasks. Our model is the same size as BERT-Base, but outperforms the much larger BERT-Large model and other more recent approaches that incorporate discourse. We also show that Conpono yields gains of 2%-6% absolute even for tasks that do not explicitly evaluate discourse: textual entailment (RTE), common sense reasoning (COPA) and reading comprehension (ReCoRD).

SenseBERT: Driving Some Sense into BERT [Website] [PDF]
Yoav Levine, Barak Lenz, Or Dagan, Ori Ram, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham 17:00–18:00

The ability to learn from large unlabeled corpora has allowed neural language models to advance the frontier in natural language understanding. However, existing self-supervision techniques operate at the word form level, which serves as a surrogate for the underlying semantic content. This paper proposes a method to employ weak-supervision directly at the word sense level. Our model, named SenseBERT, is pre-trained to predict not only the masked words but also their WordNet supersenses. Accordingly, we attain a lexical-semantic level language model, without the use of human annotation. SenseBERT achieves significantly improved lexical understanding, as we demonstrate by experimenting on SemEval Word Sense Disambiguation, and by attaining a state of the art result on the 'Word in Context' task.

[TACL] SpanBERT: Improving Pre-training by Representing and Predicting Spans [Website] [PDF]
Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy 17:00–18:00

We present SpanBERT, a pre-training method that is designed to better represent and predict spans of text. Our approach extends BERT by (1) masking contiguous random spans, rather than random tokens, and (2) training the span boundary representations to predict the entire content of the masked span, without relying on the individual token representations within it. SpanBERT consistently outperforms BERT and our better-tuned baselines, with substantial

gains on span selection tasks such as question answering and coreference resolution. In particular, with the same training data and model size as BERT-Large, our single model obtains 94.6% and 88.7% F1 on SQuAD 1.1 and 2.0 respectively. We also achieve a new state of the art on the OntoNotes coreference resolution task (79.6% F1), strong performance on the TACRED relation extraction benchmark, and even gains on GLUE.

Session 9A: Resources and Evaluation-9

A Recipe for Creating Multimodal Aligned Datasets for Sequential Tasks [Website][PDF]
Angela Lin, Sudha Rao, Asli Celikyilmaz, Elnaz Nouri, Chris Brockett, Debadeepta Dey, and Bill Dolan
17:00–18:00

Many high-level procedural tasks can be decomposed into sequences of instructions that vary in their order and choice of tools. In the cooking domain, the web offers many, partially-overlapping, text and video recipes (i.e. procedures) that describe how to make the same dish (i.e. high-level task). Aligning instructions for the same dish across different sources can yield descriptive visual explanations that are far richer semantically than conventional textual instructions, providing commonsense insight into how real-world procedures are structured. Learning to align these different instruction sets is challenging because: a) different recipes vary in their order of instructions and use of ingredients; and b) video instructions can be noisy and tend to contain far more information than text instructions. To address these challenges, we use an unsupervised alignment algorithm that learns pairwise alignments between instructions of different recipes for the same dish. We then use a graph algorithm to derive a joint alignment between multiple text and multiple video recipes for the same dish. We release the MICROSOFT RESEARCH MULTIMODAL ALIGNED RECIPE CORPUS containing ~150K pairwise alignments between recipes across 4262 dishes with rich commonsense information.

ASSET: A Dataset for Tuning and Evaluation of Sentence Simplification Models with Multiple Rewriting Transformations [Website][PDF]
Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia 17:00–18:00

In order to simplify a sentence, human editors perform multiple rewriting transformations: they split it into several shorter sentences, paraphrase words (i.e. replacing complex words or phrases by simpler synonyms), reorder components, and/or delete information deemed unnecessary. Despite these varied range of possible text alterations, current models for automatic sentence simplification are evaluated using datasets that are focused on a single transformation, such as lexical paraphrasing or splitting. This makes it impossible to understand the ability of simplification models in more realistic settings. To alleviate this limitation, this paper introduces ASSET, a new dataset for assessing sentence simplification in English. ASSET is a crowdsourced multi-reference corpus where each simplification was produced by executing several rewriting transformations. Through quantitative and qualitative experiments, we show that simplifications in ASSET are better at capturing characteristics of simplicity when compared to other standard evaluation datasets for the task. Furthermore, we motivate the need for developing better methods for automatic evaluation using ASSET, since we show that current popular metrics may not be suitable when multiple simplification transformations are performed.

Adversarial NLI: A New Benchmark for Natural Language Understanding [Website][PDF]
Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela 17:00–18:00

We introduce a new large-scale NLI benchmark dataset, collected via an iterative, adversarial human-and-model-in-the-loop procedure. We show that training models on this new dataset leads to state-of-the-art performance on a variety of popular NLI benchmarks, while posing a more difficult challenge with its new test set. Our analysis sheds light on the shortcomings of current state-of-the-art models, and shows that non-expert annotators are successful at finding their weaknesses. The data collection method can be applied in a never-ending learning scenario, becoming a moving target for NLU, rather than a static benchmark that will quickly saturate.

Beyond Accuracy: Behavioral Testing of NLP Models with CheckList [Website][PDF]
Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh 17:00–18:00

Although measuring held-out accuracy has been the primary approach to evaluate generalization, it often overestimates the performance of NLP models, while alternative approaches for evaluating models either focus on individual tasks or on specific behaviors. Inspired by principles of behavioral testing in software engineering, we introduce CheckList, a task-agnostic methodology for testing NLP models. CheckList includes a matrix of general linguistic capabilities and test types that facilitate comprehensive test ideation, as well as a software tool to generate a large and diverse number of test cases quickly. We illustrate the utility of CheckList with tests for three tasks, identifying critical failures in both commercial and state-of-art models. In a user study, a team responsible for a commercial sentiment analysis model found new and actionable bugs in an extensively tested model. In another user study, NLP practitioners with CheckList created twice as many tests, and found almost three times as many bugs as users without it.

ChartDialogs: Plotting from Natural Language Instructions [Website][PDF]
Yutong Shao and Ndapa Nakashole 17:00–18:00

This paper presents the problem of conversational plotting agents that carry out plotting actions from natural language instructions. To facilitate the development of such agents, we introduce ChartDialogs, a new multi-turn dialog dataset, covering a popular plotting library, matplotlib. The dataset contains over 15,000 dialog turns from 3,200 dialogs covering the majority of matplotlib plot types. Extensive experiments show the best-performing method achieving 61% plotting accuracy, demonstrating that the dataset presents a non-trivial challenge for future research on this task.

Code and Named Entity Recognition in StackOverflow [Website][PDF]
Jeniya Tabassum, Mounica Maddela, Wei Xu, and Alan Ritter 17:00–18:00

There is an increasing interest in studying natural language and computer code together, as large corpora of programming texts become readily available on the Internet. For example, StackOverflow currently has over 15 million

programming related questions written by 8.5 million users. Meanwhile, there is still a lack of fundamental NLP techniques for identifying code tokens or software-related named entities that appear within natural language sentences. In this paper, we introduce a new named entity recognition (NER) corpus for the computer programming domain, consisting of 15,372 sentences annotated with 20 fine-grained entity types. We trained in-domain BERT representations (BERTOverflow) on 152 million sentences from StackOverflow, which lead to an absolute increase of +10 F_1 score over off-the-shelf BERT. We also present the SoftNER model which achieves an overall 79.10 F-1 score for code and named entity recognition on StackOverflow data. Our SoftNER model incorporates a context-independent code token classifier with corpus-level features to improve the BERT-based tagging model. Our code and data are available at: https://github.com/jeniyat/StackOverflowNER/

Dialogue-Based Relation Extraction [Website][PDF]
Dian Yu, Kai Sun, Claire Cardie, and Dong Yu 17:00–18:00

We present the first human-annotated dialogue-based relation extraction (RE) dataset DialogRE, aiming to support the prediction of relation(s) between two arguments that appear in a dialogue. We further offer DialogRE as a platform for studying cross-sentence RE as most facts span multiple sentences. We argue that speaker-related information plays a critical role in the proposed task, based on an analysis of similarities and differences between dialogue-based and traditional RE tasks. Considering the timeliness of communication in a dialogue, we design a new metric to evaluate the performance of RE methods in a conversational setting and investigate the performance of several representative RE methods on DialogRE. Experimental results demonstrate that a speaker-aware extension on the best-performing model leads to gains in both the standard and conversational evaluation settings. DialogRE is available at https://dataset.org/dialogre/.

Facet-Aware Evaluation for Extractive Summarization [Website][PDF]
Yuning Mao, Liyuan Liu, Qi Zhu, Xiang Ren, and Jiawei Han 17:00–18:00

Commonly adopted metrics for extractive summarization focus on lexical overlap at the token level. In this paper, we present a facet-aware evaluation setup for better assessment of the information coverage in extracted summaries. Specifically, we treat each sentence in the reference summary as a *facet*, identify the sentences in the document that express the semantics of each facet as *support sentences* of the facet, and automatically evaluate extractive summarization methods by comparing the indices of extracted sentences and support sentences of all the facets in the reference summary. To facilitate this new evaluation setup, we construct an extractive version of the CNN/Daily Mail dataset and perform a thorough quantitative investigation, through which we demonstrate that facet-aware evaluation manifests better correlation with human judgment than ROUGE, enables fine-grained evaluation as well as comparative analysis, and reveals valuable insights of state-of-the-art summarization methods. Data can be found at `https://github.com/morningmoni/FAR`.

Fatality Killed the Cat or: BabelPic, a Multimodal Dataset for Non-Concrete Concepts [Website][PDF]
Agostina Calabrese, Michele Bevilacqua, and Roberto Navigli 17:00–18:00

Thanks to the wealth of high-quality annotated images available in popular repositories such as ImageNet, multimodal language-vision research is in full bloom. However, events, feelings and many other kinds of concepts which can be visually grounded are not well represented in current datasets. Nevertheless, we would expect a wide-coverage language understanding system to be able to classify images depicting recess and remorse, not just cats, dogs and bridges. We fill this gap by presenting BabelPic, a hand-labeled dataset built by cleaning the image-synset association found within the BabelNet Lexical Knowledge Base (LKB). BabelPic explicitly targets non-concrete concepts, thus providing refreshing new data for the community. We also show that pre-trained language-vision systems can be used to further expand the resource by exploiting natural language knowledge available in the LKB. BabelPic is available for download at http://babelpic.org.

[CL] LINSPECTOR: Multilingual Probing Tasks for Word Representations [Website][PDF]
Gözde Gül Şahin, Clara Vania, Ilia Kuznetsov, and Iryna Gurevych 17:00–18:00

Despite an ever-growing number of word representation models introduced for a large number of languages, there is a lack of a standardized technique to provide insights into what is captured by these models. Such insights would help the community to get an estimate of the downstream task performance, as well as to design more informed neural architectures, while avoiding extensive experimentation that requires substantial computational resources not all researchers have access to. A recent development in NLP is to use simple classification tasks, also called probing tasks, that test for a single linguistic feature such as part-of-speech. Existing studies mostly focus on exploring the linguistic information encoded by the continuous representations of English text. However, from a typological perspective the morphologically poor English is rather an outlier: The information encoded by the word order and function words in English is often stored on a subword, morphological level in other languages. To address this, we introduce 15 type-level probing tasks such as case marking, possession, word length, morphological tag count, and pseudoword identification for 24 languages. We present a reusable methodology for creation and evaluation of such tests in a multilingual setting, which is challenging because of a lack of resources, lower quality of tools, and differences among languages. We then present experiments on several diverse multilingual word embedding models, in which we relate the probing task performance for a diverse set of languages to a range of five classic NLP tasks: POS-tagging, dependency parsing, semantic role labeling, named entity recognition, and natural language inference. We find that a number of probing tests have significantly high positive correlation to the downstream tasks, especially for morphologically rich languages. We show that our test scan be used to explore word embeddings or black-box neural models for linguistic cues in a multilingual setting. We release the probing data sets and the evaluation suite LINSPECTOR with https://github.com/UKPLab/linspector.

More Diverse Dialogue Datasets via Diversity-Informed Data Collection [Website][PDF]
Katherine Stasaski, Grace Hui Yang, and Marti A. Hearst 17:00–18:00

Automated generation of conversational dialogue using modern neural architectures has made notable advances. However, these models are known to have a drawback of often producing uninteresting, predictable responses; this is known as the diversity problem. We introduce a new strategy to address this problem, called Diversity-Informed Data Collection. Unlike prior approaches, which modify model architectures to solve the problem, this method uses dynamically computed corpus-level statistics to determine which conversational participants to collect data from. Diversity-Informed Data Collection produces significantly more diverse data than baseline data collection methods, and better results on two downstream tasks: emotion classification and dialogue generation. This method is generalizable and can be used with other corpus-level metrics.

ParaCrawl: Web-Scale Acquisition of Parallel Corpora [Website] [PDF]
Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza 17:00–18:00

We report on methods to create the largest publicly available parallel corpora by crawling the web, using open source software. We empirically compare alternative methods and publish benchmark data sets for sentence alignment and sentence pair filtering. We also describe the parallel corpora released and evaluate their quality and their usefulness to create machine translation systems.

S2ORC: The Semantic Scholar Open Research Corpus [Website] [PDF]
Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld 17:00–18:00

We introduce S2ORC, a large corpus of 81.1M English-language academic papers spanning many academic disciplines. The corpus consists of rich metadata, paper abstracts, resolved bibliographic references, as well as structured full text for 8.1M open access papers. Full text is annotated with automatically-detected inline mentions of citations, figures, and tables, each linked to their corresponding paper objects. In S2ORC, we aggregate papers from hundreds of academic publishers and digital archives into a unified source, and create the largest publicly-available collection of machine-readable academic text to date. We hope this resource will facilitate research and development of tools and tasks for text mining over academic text.

Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics [Website] [PDF]
Nitika Mathur, Timothy Baldwin, and Trevor Cohn 17:00–18:00

Automatic metrics are fundamental for the development and evaluation of machine translation systems. Judging whether, and to what extent, automatic metrics concur with the gold standard of human evaluation is not a straightforward problem. We show that current methods for judging metrics are highly sensitive to the translations used for assessment, particularly the presence of outliers, which often leads to falsely confident conclusions about a metric's efficacy. Finally, we turn to pairwise system ranking, developing a method for thresholding performance improvement under an automatic metric against human judgements, which allows quantification of type I versus type II errors incurred, i.e., insignificant human differences in system quality that are accepted, and significant human differences that are rejected. Together, these findings suggest improvements to the protocols for metric evaluation and system performance evaluation in machine translation.

Session 9A: Student Research Workshop

Checkpoint Reranking: An Approach to Select Better Hypothesis for Neural Machine Translation Systems [Website][PDF]
Vinay Pandramish and Dipti Misra Sharma 17:00–18:00

In this paper, we propose a method of re-ranking the outputs of Neural Machine Translation (NMT) systems. After the decoding process, we select a few last iteration outputs in the training process as the N-best list. After training a Neural Machine Translation (NMT) baseline system, it has been observed that these iteration outputs have an oracle score higher than baseline up to 1.01 BLEU points compared to the last iteration of the trained system.We come up with a ranking mechanism by solely focusing on the decoder's ability to generate distinct tokens and without the usage of any language model or data. With this method, we achieved a translation improvement up to +0.16 BLEU points over baseline.We also evaluate our approach by applying the coverage penalty to the training process.In cases of moderate coverage penalty, the oracle scores are higher than the final iteration up to +0.99 BLEU points, and our algorithm gives an improvement up to +0.17 BLEU points.With excessive penalty, there is a decrease in translation quality compared to the baseline system. Still, an increase in oracle scores up to +1.30 is observed with the re-ranking algorithm giving an improvement up to +0.15 BLEU points is found in case of excessive penalty.The proposed re-ranking method is a generic one and can be extended to other language pairs as well.

Cross-Lingual Disaster-related Multi-label Tweet Classification with Manifold Mixup [Website][PDF]
Jishnu Ray Chowdhury, Cornelia Caragea, and Doina Caragea 17:00–18:00

Distinguishing informative and actionable messages from a social media platform like Twitter is critical for facilitating disaster management. For this purpose, we compile a multilingual dataset of over 130K samples for multi-label classification of disaster-related tweets. We present a masking-based loss function for partially labelled samples and demonstrate the effectiveness of Manifold Mixup in the text domain. Our main model is based on Multilingual BERT, which we further improve with Manifold Mixup. We show that our model generalizes to unseen disasters in the test set. Furthermore, we analyze the capability of our model for zero-shot generalization to new languages. Our code, dataset, and other resources are available on Github.

Inducing Grammar from Long Short-Term Memory Networks by Shapley Decomposition [Website][PDF]
Yuhui Zhang and Allen Nie 17:00–18:00

The principle of compositionality has deep roots in linguistics: the meaning of an expression is determined by its structure and the meanings of its constituents. However, modern neural network models such as long short-term memory network process expressions in a linear fashion and do not seem to incorporate more complex compositional patterns. In this work, we show that we can explicitly induce grammar by tracing the computational process of a long short-term memory network. We show: (i) the multiplicative nature of long short-term memory network allows complex interaction beyond sequential linear combination; (ii) we can generate compositional trees from the network without external linguistic knowledge; (iii) we evaluate the syntactic difference between the generated trees, randomly generated trees and gold reference trees produced by constituency parsers; (iv) we evaluate whether the generated trees contain the rich semantic information.

Exploring the Role of Context to Distinguish Rhetorical and Information-Seeking Questions [Website][PDF]
Yuan Zhuang and Ellen Riloff 17:00–18:00

Social media posts often contain questions, but many of the questions are rhetorical and do not seek information. Our work studies the problem of distinguishing rhetorical and information-seeking questions on Twitter. Most work has focused on features of the question itself, but we hypothesize that the prior context plays a role too. This paper introduces a new dataset containing questions in tweets paired with their prior tweets to provide context. We create classification models to assess the difficulty of distinguishing rhetorical and information-seeking questions, and experiment with different properties of the prior context. Our results show that the prior tweet and topic features can improve performance on this task.

Session 9A: Summarization-4

A Transformer-based Approach for Source Code Summarization [Website][PDF]
Wasi Ahmad, Saikat Chakraborty, Baishakhi Ray, and Kai-Wei Chang 17:00–18:00

Generating a readable summary that describes the functionality of a program is known as source code summarization. In this task, learning code representation by modeling the pairwise relationship between code tokens to capture their long-range dependencies is crucial. To learn code representation for summarization, we explore the Transformer model that uses a self-attention mechanism and has shown to be effective in capturing long-range dependencies. In this work, we show that despite the approach is simple, it outperforms the state-of-the-art techniques by a significant margin. We perform extensive analysis and ablation studies that reveal several important findings, e.g., the absolute encoding of source code tokens' position hinders, while relative encoding significantly improves the summarization performance. We have made our code publicly available[3] to facilitate future research.

Asking and Answering Questions to Evaluate the Factual Consistency of Summaries [Website][PDF]
Alex Wang, Kyunghyun Cho, and Mike Lewis 17:00–18:00

Practical applications of abstractive summarization models are limited by frequent factual inconsistencies with respect to their input. Existing automatic evaluation metrics for summarization are largely insensitive to such errors. We propose QAGS (pronounced "kags"), an automatic evaluation protocol that is designed to identify factual inconsistencies in a generated summary. QAGS is based on the intuition that if we ask questions about a summary and its source, we will receive similar answers if the summary is factually consistent with the source. To evaluate QAGS, we collect human judgments of factual consistency on model-generated summaries for the CNN/DailyMail (Hermann et al., 2015) and XSUM (Narayan et al., 2018) summarization datasets. QAGS has substantially higher correlations with these judgments than other automatic evaluation metrics. Also, QAGS offers a natural form of interpretability: The answers and questions generated while computing QAGS indicate which tokens of a summary are inconsistent and why. We believe QAGS is a promising tool in automatically generating usable and factually consistent text. Code for QAGS will be available at https://github.com/W4ngatang/qags.

Discourse-Aware Neural Extractive Text Summarization [Website][PDF]
Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu 17:00–18:00

Recently BERT has been adopted for document encoding in state-of-the-art text summarization models. However, sentence-based extractive models often result in redundant or uninformative phrases in the extracted summaries. Also, long-range dependencies throughout a document are not well captured by BERT, which is pre-trained on sentence pairs instead of documents. To address these issues, we present a discourse-aware neural summarization model - DiscoBert. DiscoBert extracts sub-sentential discourse units (instead of sentences) as candidates for extractive selection on a finer granularity. To capture the long-range dependencies among discourse units, structural discourse graphs are constructed based on RST trees and coreference mentions, encoded with Graph Convolutional Networks. Experiments show that the proposed model outperforms state-of-the-art methods by a significant margin on popular summarization benchmarks compared to other BERT-base models.

Discrete Optimization for Unsupervised Sentence Summarization with Word-Level Extraction [Website][PDF]
Raphael Schumann, Lili Mou, Yao Lu, Olga Vechtomova, and Katja Markert 17:00–18:00

Automatic sentence summarization produces a shorter version of a sentence, while preserving its most important information. A good summary is characterized by language fluency and high information overlap with the source sentence. We model these two aspects in an unsupervised objective function, consisting of language modeling and semantic similarity metrics. We search for a high-scoring summary by discrete optimization. Our proposed method achieves a new state-of-the art for unsupervised sentence summarization according to ROUGE scores. Additionally, we demonstrate that the commonly reported ROUGE F1 metric is sensitive to summary length. Since this is unwillingly exploited in recent work, we emphasize that future evaluation should explicitly group summarization systems by output length brackets.

Exploring Content Selection in Summarization of Novel Chapters [Website][PDF]
Faisal Ladhak, Bryan Li, Yaser Al-Onaizan, and Kathleen McKeown 17:00–18:00

We present a new summarization task, generating summaries of novel chapters using summary/chapter pairs from online study guides. This is a harder task than the news summarization task, given the chapter length as well as the extreme paraphrasing and generalization found in the summaries. We focus on extractive summarization, which requires the creation of a gold-standard set of extractive summaries. We present a new metric for aligning reference summary sentences with chapter sentences to create gold extracts and also experiment with different alignment methods. Our experiments demonstrate significant improvement over prior alignment approaches for our task as shown through automatic metrics and a crowd-sourced pyramid analysis.

FEQA: A Question Answering Evaluation Framework for Faithfulness Assessment in Abstractive Summarization [Website][PDF]
Esin Durmus, He He, and Mona Diab 17:00–18:00

Neural abstractive summarization models are prone to generate content inconsistent with the source document, i.e. unfaithful. Existing automatic metrics do not capture such mistakes effectively. We tackle the problem of evaluating

[3] https://github.com/wasiahmad/NeuralCodeSum

faithfulness of a generated summary given its source document. We first collected human annotations of faithfulness for outputs from numerous models on two datasets. We find that current models exhibit a trade-off between abstractiveness and faithfulness: outputs with less word overlap with the source document are more likely to be unfaithful. Next, we propose an automatic question answering (QA) based metric for faithfulness, FEQA, which leverages recent advances in reading comprehension. Given question-answer pairs generated from the summary, a QA model extracts answers from the document; non-matched answers indicate unfaithful information in the summary. Among metrics based on word overlap, embedding similarity, and learned language understanding models, our QA-based metric has significantly higher correlation with human faithfulness scores, especially on highly abstractive summaries.

Fact-based Content Weighting for Evaluating Abstractive Summarisation [Website][PDF]
Xinnuo Xu, Ondřej Dušek, Jingyi Li, Verena Rieser, and Ioannis Konstas 17:00–18:00

Abstractive summarisation is notoriously hard to evaluate since standard word-overlap-based metrics are insufficient. We introduce a new evaluation metric which is based on fact-level content weighting, i.e. relating the facts of the document to the facts of the summary. We fol- low the assumption that a good summary will reflect all relevant facts, i.e. the ones present in the ground truth (human-generated refer- ence summary). We confirm this hypothe- sis by showing that our weightings are highly correlated to human perception and compare favourably to the recent manual highlight- based metric of Hardy et al. (2019).

Hooks in the Headline: Learning to Generate Headlines with Controlled Styles [Website][PDF]
Di Jin, Zhijing Jin, Joey Tianyi Zhou, Lisa Orii, and Peter Szolovits 17:00–18:00

Current summarization systems only produce plain, factual headlines, far from the practical needs for the exposure and memorableness of the articles. We propose a new task, Stylistic Headline Generation (SHG), to enrich the headlines with three style options (humor, romance and clickbait), thus attracting more readers. With no style-specific article-headline pair (only a standard headline summarization dataset and mono-style corpora), our method TitleStylist generates stylistic headlines by combining the summarization and reconstruction tasks into a multitasking framework. We also introduced a novel parameter sharing scheme to further disentangle the style from text. Through both automatic and human evaluation, we demonstrate that TitleStylist can generate relevant, fluent headlines with three target styles: humor, romance, and clickbait. The attraction score of our model generated headlines outperforms the state-of-the-art summarization model by 9.68%, even outperforming human-written references.

Knowledge Graph-Augmented Abstractive Summarization with Semantic-Driven Cloze Reward [Website][PDF]
Luyang Huang, Lingfei Wu, and Lu Wang 17:00–18:00

Sequence-to-sequence models for abstractive summarization have been studied extensively, yet the generated summaries commonly suffer from fabricated content, and are often found to be near-extractive. We argue that, to address these issues, the summarizer should acquire semantic interpretation over input, e.g., via structured representation, to allow the generation of more informative summaries. In this paper, we present ASGARD, a novel framework for Abstractive Summarization with Graph-Augmentation and semantic-driven RewarD. We propose the use of dual encoders—a sequential document encoder and a graph-structured encoder—to maintain the global context and local characteristics of entities, complementing each other. We further design a reward based on a multiple choice cloze test to drive the model to better capture entity interactions. Results show that our models produce significantly higher ROUGE scores than a variant without knowledge graph as input on both New York Times and CNN/Daily Mail datasets. We also obtain better or comparable performance compared to systems that are fine-tuned from large pretrained language models. Human judges further rate our model outputs as more informative and containing fewer unfaithful errors.

Optimizing the Factual Correctness of a Summary: A Study of Summarizing Radiology Reports [Website][PDF]
Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D. Manning, and Curtis Langlotz 17:00–18:00

Neural abstractive summarization models are able to generate summaries which have high overlap with human references. However, existing models are not optimized for factual correctness, a critical metric in real-world applications. In this work, we develop a general framework where we evaluate the factual correctness of a generated summary by fact-checking it automatically against its reference using an information extraction module. We further propose a training strategy which optimizes a neural summarization model with a factual correctness reward via reinforcement learning. We apply the proposed method to the summarization of radiology reports, where factual correctness is a key requirement. On two separate datasets collected from hospitals, we show via both automatic and human evaluation that the proposed approach substantially improves the factual correctness and overall quality of outputs over a competitive neural summarization system, producing radiology summaries that approach the quality of human-authored ones.

Storytelling with Dialogue: A Critical Role Dungeons and Dragons Dataset [Website][PDF]
Revanth Rameshkumar and Peter Bailey 17:00–18:00

This paper describes the Critical Role Dungeons and Dragons Dataset (CRD3) and related analyses. Critical Role is an unscripted, live-streamed show where a fixed group of people play Dungeons and Dragons, an open-ended role-playing game. The dataset is collected from 159 Critical Role episodes transcribed to text dialogues, consisting of 398,682 turns. It also includes corresponding abstractive summaries collected from the Fandom wiki. The dataset is linguistically unique in that the narratives are generated entirely through player collaboration and spoken interaction. For each dialogue, there are a large number of turns, multiple abstractive summaries with varying levels of detail, and semantic ties to the previous dialogues. In addition, we provide a data augmentation method that produces 34,243 summary-dialogue chunk pairs to support current neural ML approaches, and we provide an abstractive summariza-

tion benchmark and evaluation.

The Summary Loop: Learning to Write Abstractive Summaries Without Examples [Website][PDF]
Philippe Laban, Andrew Hsi, John Canny, and Marti A. Hearst 17:00–18:00

This work presents a new approach to unsupervised abstractive summarization based on maximizing a combination of coverage and fluency for a given length constraint. It introduces a novel method that encourages the inclusion of key terms from the original document into the summary: key terms are masked out of the original document and must be filled in by a coverage model using the current generated summary. A novel unsupervised training procedure leverages this coverage model along with a fluency model to generate and score summaries. When tested on popular news summarization datasets, the method outperforms previous unsupervised methods by more than 2 R-1 points, and approaches results of competitive supervised methods. Our model attains higher levels of abstraction with copied passages roughly two times shorter than prior work, and learns to compress and merge sentences without supervision.

Unsupervised Opinion Summarization as Copycat-Review Generation [Website][PDF]
Arthur Bražinskas, Mirella Lapata, and Ivan Titov 17:00–18:00

Opinion summarization is the task of automatically creating summaries that reflect subjective information expressed in multiple documents, such as product reviews. While the majority of previous work has focused on the extractive setting, i.e., selecting fragments from input reviews to produce a summary, we let the model generate novel sentences and hence produce abstractive summaries. Recent progress in summarization has seen the development of supervised models which rely on large quantities of document-summary pairs. Since such training data is expensive to acquire, we instead consider the unsupervised setting, in other words, we do not use any summaries in training. We define a generative model for a review collection which capitalizes on the intuition that when generating a new review given a set of other reviews of a product, we should be able to control the "amount of novelty" going into the new review or, equivalently, vary the extent to which it deviates from the input. At test time, when generating summaries, we force the novelty to be minimal, and produce a text reflecting consensus opinions. We capture this intuition by defining a hierarchical variational autoencoder model. Both individual reviews and the products they correspond to are associated with stochastic latent codes, and the review generator ("decoder") has direct access to the text of input reviews through the pointer-generator mechanism. Experiments on Amazon and Yelp datasets, show that setting at test time the review's latent code to its mean, allows the model to produce fluent and coherent summaries reflecting common opinions.

Session 9A: Theme-1

(Re)construing Meaning in NLP [Website][PDF]
Sean Trott, Tiago Timponi Torrent, Nancy Chang, and Nathan Schneider 17:00–18:00

Human speakers have an extensive toolkit of ways to express themselves. In this paper, we engage with an idea largely absent from discussions of meaning in natural language understanding—namely, that the way something is expressed reflects different ways of conceptualizing or construing the information being conveyed. We first define this phenomenon more precisely, drawing on considerable prior work in theoretical cognitive semantics and psycholinguistics. We then survey some dimensions of construed meaning and show how insights from construal could inform theoretical and practical work in NLP.

Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data [Website][PDF]
Emily M. Bender and Alexander Koller 17:00–18:00

The success of the large neural language models on many NLP tasks is exciting. However, we find that these successes sometimes lead to hype in which these models are being described as "understanding" language or capturing "meaning". In this position paper, we argue that a system trained only on form has a priori no way to learn meaning. In keeping with the ACL 2020 theme of "Taking Stock of Where We've Been and Where We're Going", we argue that a clear understanding of the distinction between form and meaning will help guide the field towards better science around natural language understanding.

Examining Citations of Natural Language Processing Literature [Website][PDF]
Saif M. Mohammad 17:00–18:00

We extracted information from the ACL Anthology (AA) and Google Scholar (GS) to examine trends in citations of NLP papers. We explore questions such as: how well cited are papers of different types (journal articles, conference papers, demo papers, etc.)? how well cited are papers from different areas of within NLP? etc. Notably, we show that only about 56% of the papers in AA are cited ten or more times. CL Journal has the most cited papers, but its citation dominance has lessened in recent years. On average, long papers get almost three times as many citations as short papers; and papers on sentiment classification, anaphora resolution, and entity recognition have the highest median citations. The analyses presented here, and the associated dataset of NLP papers mapped to citations, have a number of uses including: understanding how the field is growing and quantifying the impact of different types of papers.

How Can We Accelerate Progress Towards Human-like Linguistic Generalization? [Website][PDF]
Tal Linzen 17:00–18:00

This position paper describes and critiques the Pretraining-Agnostic Identically Distributed (PAID) evaluation paradigm, which has become a central tool for measuring progress in natural language understanding. This paradigm consists of three stages: (1) pre-training of a word prediction model on a corpus of arbitrary size; (2) fine-tuning (transfer learning) on a training set representing a classification task; (3) evaluation on a test set drawn from the same distribution as that training set. This paradigm favors simple, low-bias architectures, which, first, can be scaled to process vast amounts of data, and second, can capture the fine-grained statistical properties of a particular data set, regardless of whether those properties are likely to generalize to examples of the task outside the data set. This contrasts with humans, who learn language from several orders of magnitude less data than the systems favored by this evaluation paradigm, and generalize to new tasks in a consistent way. We advocate for supplementing or replacing PAID with paradigms that reward architectures that generalize as quickly and robustly as humans.

How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence [Website][PDF]
Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun 17:00–18:00

Legal Artificial Intelligence (LegalAI) focuses on applying the technology of artificial intelligence, especially natural language processing, to benefit tasks in the legal domain. In recent years, LegalAI has drawn increasing attention rapidly from both AI researchers and legal professionals, as LegalAI is beneficial to the legal system for liberating legal professionals from a maze of paperwork. Legal professionals often think about how to solve tasks from rule-based and symbol-based methods, while NLP researchers concentrate more on data-driven and embedding methods. In this paper, we introduce the history, the current state, and the future directions of research in LegalAI. We illustrate the tasks from the perspectives of legal professionals and NLP researchers and show several representative applications in LegalAI. We conduct experiments and provide an in-depth analysis of the advantages and disadvantages of existing works to explore possible future directions. You can find the implementation of our work from https://github.com/thunlp/CLAIM.

Intermediate-Task Transfer Learning with Pretrained Language Models: When and Why Does It Work? [Website][PDF]
Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman 17:00–18:00

While pretrained models such as BERT have shown large gains across natural language understanding tasks, their performance can be improved by further training the model on a data-rich intermediate task, before fine-tuning it on a target task. However, it is still poorly understood when and why intermediate-task training is beneficial for a given target task. To investigate this, we perform a large-scale study on the pretrained RoBERTa model with 110 intermediate-target task combinations. We further evaluate all trained models with 25 probing tasks meant to reveal the specific skills that drive transfer. We observe that intermediate tasks requiring high-level inference and reasoning abilities tend to work best. We also observe that target task performance is strongly correlated with higher-level

abilities such as coreference resolution. However, we fail to observe more granular correlations between probing and target task performance, highlighting the need for further work on broad-coverage probing benchmarks. We also observe evidence that the forgetting of knowledge learned during pretraining may limit our analysis, highlighting the need for further work on transfer learning methods in these settings.

Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview [Website] [PDF]
Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy 17:00–18:00

An increasing number of natural language processing papers address the effect of bias on predictions, introducing mitigation techniques at different parts of the standard NLP pipeline (data and models). However, these works have been conducted individually, without a unifying framework to organize efforts within the field. This situation leads to repetitive approaches, and focuses overly on bias symptoms/effects, rather than on their origins, which could limit the development of effective countermeasures. In this paper, we propose a unifying predictive bias framework for NLP. We summarize the NLP literature and suggest general mathematical definitions of predictive bias. We differentiate two consequences of bias: outcome disparities and error disparities, as well as four potential origins of biases: label bias, selection bias, model overamplification, and semantic bias. Our framework serves as an overview of predictive bias in NLP, integrating existing work into a single structure, and providing a conceptual baseline for improved frameworks.

What Does BERT with Vision Look At? [Website] [PDF]
Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang 17:00–18:00

Pre-trained visually grounded language models such as ViLBERT, LXMERT, and UNITER have achieved significant performance improvement on vision-and-language tasks but what they learn during pre-training remains unclear. In this work, we demonstrate that certain attention heads of a visually grounded language model actively ground elements of language to image regions. Specifically, some heads can map entities to image regions, performing the task known as entity grounding. Some heads can even detect the syntactic relations between non-entity words and image regions, tracking, for example, associations between verbs and regions corresponding to their arguments. We denote this ability as *syntactic grounding*. We verify grounding both quantitatively and qualitatively, using Flickr30K Entities as a testbed.

Demo Session 4B

Time: *17:45–18:30*

BENTO: A Visual Platform for Building Clinical NLP Pipelines Based on CodaLab [Website][PDF]
Yonghao Jin, Fei Li, and Hong Yu

CodaLab is an open-source web-based platform for collaborative computational research. Although CodaLab has gained popularity in the research community, its interface has limited support for creating reusable tools that can be easily applied to new datasets and composed into pipelines. In clinical domain, natural language processing (NLP) on medical notes generally involves multiple steps, like tokenization, named entity recognition, etc. Since these steps require different tools which are usually scattered in different publications, it is not easy for researchers to use them to process their own datasets. In this paper, we present BENTO, a workflow management platform with a graphic user interface (GUI) that is built on top of CodaLab, to facilitate the process of building clinical NLP pipelines. BENTO comes with a number of clinical NLP tools that have been pre-trained using medical notes and expert annotations and can be readily used for various clinical NLP tasks. It also allows researchers and developers to create their custom tools (e.g., pre-trained NLP models) and use them in a controlled and reproducible way. In addition, the GUI interface enables researchers with limited computer background to compose tools into NLP pipelines and then apply the pipelines on their own datasets in a "what you see is what you get" (WYSIWYG) way. Although BENTO is designed for clinical NLP applications, the underlying architecture is flexible to be tailored to any other domains.

Interactive Task Learning from GUI-Grounded Natural Language Instructions and Demonstrations [Website][PDF]
Toby Jia-Jun Li, Tom Mitchell, and Brad Myers

We show SUGILITE, an intelligent task automation agent that can learn new tasks and relevant associated concepts interactively from the user's natural language instructions and demonstrations, using the graphical user interfaces (GUIs) of third-party mobile apps. This system provides several interesting features: (1) it allows users to teach new task procedures and concepts through verbal instructions together with demonstration of the steps of a script using GUIs; (2) it supports users in clarifying their intents for demonstrated actions using GUI-grounded verbal instructions; (3) it infers parameters of tasks and their possible values in utterances using the hierarchical structures of the underlying app GUIs; and (4) it generalizes taught concepts to different contexts and task domains. We describe the architecture of the SUGILITE system, explain the design and implementation of its key features, and show a prototype in the form of a conversational assistant on Android.

MixingBoard: a Knowledgeable Stylized Integrated Text Generation Platform [Website][PDF]
Xiang Gao, Michel Galley, and Bill Dolan

We present MixingBoard, a platform for quickly building demos with a focus on knowledge grounded stylized text generation. We unify existing text generation algorithms in a shared codebase and further adapt earlier algorithms for constrained generation. To borrow advantages from different models, we implement strategies for cross-model integration, from the token probability level to the latent space level. An interface to external knowledge is provided via a module that retrieves, on-the-fly, relevant knowledge from passages on the web or a document collection. A user interface for local development, remote webpage access, and a RESTful API are provided to make it simple for users to build their own demos.

Session 9B Overview – Tuesday, July 7, 2020 18:00–19:00

Track A *Computational Social Science and Social Media-7* Abstracts	Analyzing Political Parody in Social Media *Maronikolakis, Sánchez Villegas, Preotiuc-Pietro, and Aletras* [Website][PDF]	Balancing Objectives in Counseling Conversations: Advancing Forwards or Looking Backwards *Zhang and Danescu-Niculescu-Mizil* [Website][PDF]	Detecting Perceived Emotions in Hurricane Disasters *Desai, Caragea, and Li* [Website][PDF]	Hierarchical Modeling for User Personality Prediction: The Role of Message-Level Attention *Lynn, Balasubramanian, and Schwartz* [Website][PDF]	Measuring Forecasting Skill from Text *Zong, Ritter, and Hovy* [Website][PDF]
	Text and Causal Inference: A Review of Using Text to Remove Confounding from Causal Estimates *Keith, Jensen, and O'Connor* [Website][PDF]	Text-Based Ideal Points *Vafa, Naidu, and Blei* [Website][PDF]	Understanding the Language of Political Agreement and Disagreement in Legislative Texts *Davoodi, Waltenburg, and Goldwasser* [Website][PDF]	Would you Rather? A New Benchmark for Learning Machine Alignment with Cultural Values and Social Preferences *Tay, Ong, Fu, Chan, Chen, Luu, and Pal* [Website][PDF]	
Track B *Discourse and Pragmatics-4* Abstracts	Discourse as a Function of Event: Profiling Discourse Structure in News Articles around the Main Event *Choubey, Lee, Huang, and Wang* [Website][PDF]	Harnessing the linguistic signal to predict scalar inferences *Schuster, Chen, and Degen* [Website][PDF]	Implicit Discourse Relation Classification: We Need to Talk about Evaluation *Kim, Feng, Gunasekara, and Lastras* [Website][PDF]	PeTra: A Sparsely Supervised Memory Model for People Tracking *Toshniwal, Ettinger, Gimpel, and Livescu* [Website][PDF]	ZPR2: Joint Zero Pronoun Recovery and Resolution using Multi-Task Learning and BERT *Song, Xu, Zhang, Chen, and Yu* [Website][PDF]
Track C *Ethics and NLP-4* Abstracts	Contextualizing Hate Speech Classifiers with Post-hoc Explanation *Kennedy, Jin, Mostafazadeh Davani, Dehghani, and Ren* [Website][PDF]	Double-Hard Debias: Tailoring Word Embeddings for Gender Bias Mitigation *Wang, Lin, Rajani, McCann, Ordonez, and Xiong* [Website][PDF]	Language (Technology) is Power: A Critical Survey of "Bias" in NLP *Blodgett, Barocas, Daumé III, and Wallach* [Website][PDF]	Social Bias Frames: Reasoning about Social and Power Implications of Language *Sap, Gabriel, Qin, Jurafsky, Smith, and Choi* [Website][PDF]	Social Biases in NLP Models as Barriers for Persons with Disabilities *Hutchinson, Prabhakaran, Denton, Webster, Zhong, and Denuyl* [Website][PDF]
	Towards Debiasing Sentence Representations *Liang, Li, Zheng, Lim, Salakhutdinov, and Morency* [Website][PDF]				
Track D *Interpretability and Analysis of Models for NLP-6* Abstracts	A Re-evaluation of Knowledge Graph Completion Methods *Sun, Vashishth, Sanyal, Talukdar, and Yang* [Website][PDF]	Cross-Linguistic Syntactic Evaluation of Word Prediction Models *Mueller, Nicolai, Petrou-Zeniou, Talmina, and Linzen* [Website][PDF]	[TACL] Does Syntax Need to Grow on Trees? Sources of Hierarchical Inductive Bias in Sequence-to-Sequence Networks *McCoy, Frank, and Linzen* [Website][PDF]	Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior? *Hase and Bansal* [Website][PDF]	Explaining Black Box Predictions and Unveiling Data Artifacts through Influence Functions *Han, Wallace, and Tsvetkov* [Website][PDF]

	Finding Universal Grammatical Relations in Multilingual BERT *Chi, Hewitt, and Manning* [Website][PDF]	Generating Hierarchical Explanations on Text Classification via Feature Interaction Detection *Chen, Zheng, and Ji* [Website][PDF]	Obtaining Faithful Interpretations from Compositional Neural Networks *Subramanian, Bogin, Gupta, Wolfson, Singh, Berant, and Gardner* [Website][PDF]	On the Cross-lingual Transferability of Monolingual Representations *Artetxe, Ruder, and Yogatama* [Website][PDF]	Rationalizing Text Matching: Learning Sparse Alignments via Optimal Transport *Swanson, Yu, and Lei* [Website][PDF]
Track E *Question Answering-6* Abstracts	Benefits of Intermediate Annotations in Reading Comprehension *Dua, Singh, and Gardner* [Website][PDF]	[TACL] Break It Down: A Question Understanding Benchmark *Wolfson, Geva, Gupta, Goldberg, Gardner, Deutch, and Berant* [Website][PDF]	Crossing Variational Autoencoders for Answer Retrieval *Yu, Wu, Zeng, Tao, Deng, and Jiang* [Website][PDF]	Improving Multi-hop Question Answering over Knowledge Graphs using Knowledge Base Embeddings *Saxena, Tripathi, and Talukdar* [Website][PDF]	[TACL] Investigating Prior Knowledge for Challenging Chinese Machine Reading Comprehension *Sun, Yu, Yu, and Cardie* [Website][PDF]
	Logic-Guided Data Augmentation and Regularization for Consistent Question Answering *Asai and Hajishirzi* [Website][PDF]	On the Importance of Diversity in Question Generation for QA *Sultan, Chandel, Fernandez Astudillo, and Castelli* [Website][PDF]	Probabilistic Assumptions Matter: Improved Models for Distantly-Supervised Document-Level Question Answering *Cheng, Chang, Lee, and Toutanova* [Website][PDF]	SCDE: Sentence Cloze Dataset with High Quality Distractors From Examinations *Kong, Gangal, and Hovy* [Website][PDF]	Selective Question Answering under Domain Shift *Kamath, Jia, and Liang* [Website][PDF]
	Template-Based Question Generation from Retrieved Sentences for Improved Unsupervised Question Answering *Fabbri, Ng, Wang, Nallapati, and Xiang* [Website][PDF]	The Cascade Transformer: an Application for Efficient Answer Sentence Selection *Soldaini and Moschitti* [Website][PDF]	Transformers to Learn Hierarchical Contexts in Multiparty Dialogue for Span-based Question Answering *Li and Choi* [Website][PDF]	[TACL] TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages *Clark, Palomaki, Nikolaev, Choi, Garrette, Collins, and Kwiatkowski* [Website][PDF]	
Track F *Resources and Evaluation-10* Abstracts	A Corpus for Large-Scale Phonetic Typology *Salesky, Chodroff, Pimentel, Wiesner, Cotterell, Black, and Eisner* [Website][PDF]	An Effectiveness Metric for Ordinal Classification: Formal Properties and Experimental Results *Amigo, Gonzalo, Mizzaro, and Carrillo-de-Albornoz* [Website][PDF]	Not All Claims are Created Equal: Choosing the Right Statistical Approach to Assess Hypotheses *Sadeqi Azer, Khashabi, Sabharwal, and Roth* [Website][PDF]	STARC: Structured Annotations for Reading Comprehension *Berzak, Malmaud, and Levy* [Website][PDF]	WinoWhy: A Deep Diagnosis of Essential Commonsense Knowledge for Answering Winograd Schema Challenge *Zhang, Zhao, and Song* [Website][PDF]
Track G *Sentiment Analysis, Stylistic Analysis, and Argument Mining-7* Abstracts	Agreement Prediction of Arguments in Cyber Argumentation for Detecting Stance Polarity and Intensity *Sirrianni, Liu, and Adams* [Website][PDF]	Cross-Lingual Unsupervised Sentiment Classification with Multi-View Transfer Learning *Fei and Li* [Website][PDF]	Efficient Pairwise Annotation of Argument Quality *Gienapp, Stein, Hagen, and Potthast* [Website][PDF]	Entity-Aware Dependency-Based Deep Graph Attention Network for Comparative Preference Classification *Ma, Mazumder, Wang, and Liu* [Website][PDF]	GoEmotions: A Dataset of Fine-Grained Emotions *Demszky, Movshovitz-Attias, Ko, Cowen, Nemade, and Ravi* [Website][PDF]

	OpinionDigest: A Simple Framework for Opinion Summarization *Suhara, Wang, Angelidis, and Tan* [Website][PDF]	A Comprehensive Analysis of Preprocessing for Word Representation Learning in Affective Tasks *Babanejad, Agrawal, An, and Papagelis* [Website][PDF]			
Track H *Student Research Workshop* Abstracts	Compositional Generalization by Factorizing Alignment and Translation *Russin, Jo, O'Reilly, and Bengio* [Website][PDF]	RPD: A Distance Function Between Word Embeddings *Zhou, Huang, and Zheng* [Website][PDF]	#NotAWhore! A Computational Linguistic Perspective of Rape Culture and Victimization on Social Media *Suvarna and Bhalla* [Website][PDF]	Research Replication Prediction Using Weakly Supervised Learning *Luo, Li, Wang, and Liu* [Website]	Inducing Grammar from Long Short-Term Memory Networks by Shapley Decomposition *Zhang and Nie* [Website][PDF]

Session 9B Details

Session 9B: Computational Social Science and Social Media-7

Analyzing Political Parody in Social Media [Website][PDF]
Antonios Maronikolakis, Danae Sánchez Villegas, Daniel Preotiuc-Pietro, and Nikolaos Aletras 18:00–19:00

Parody is a figurative device used to imitate an entity for comedic or critical purposes and represents a widespread phenomenon in social media through many popular parody accounts. In this paper, we present the first computational study of parody. We introduce a new publicly available data set of tweets from real politicians and their corresponding parody accounts. We run a battery of supervised machine learning models for automatically detecting parody tweets with an emphasis on robustness by testing on tweets from accounts unseen in training, across different genders and across countries. Our results show that political parody tweets can be predicted with an accuracy up to 90%. Finally, we identify the markers of parody through a linguistic analysis. Beyond research in linguistics and political communication, accurately and automatically detecting parody is important to improving fact checking for journalists and analytics such as sentiment analysis through filtering out parodical utterances.

Balancing Objectives in Counseling Conversations: Advancing Forwards or Looking Backwards [Website][PDF]
Justine Zhang and Cristian Danescu-Niculescu-Mizil 18:00–19:00

Throughout a conversation, participants make choices that can orient the flow of the interaction. Such choices are particularly salient in the consequential domain of crisis counseling, where a difficulty for counselors is balancing between two key objectives: advancing the conversation towards a resolution, and empathetically addressing the crisis situation. In this work, we develop an unsupervised methodology to quantify how counselors manage this balance. Our main intuition is that if an utterance can only receive a narrow range of appropriate replies, then its likely aim is to advance the conversation forwards, towards a target within that range. Likewise, an utterance that can only appropriately follow a narrow range of possible utterances is likely aimed backwards at addressing a specific situation within that range. By applying this intuition, we can map each utterance to a continuous orientation axis that captures the degree to which it is intended to direct the flow of the conversation forwards or backwards. This unsupervised method allows us to characterize counselor behaviors in a large dataset of crisis counseling conversations, where we show that known counseling strategies intuitively align with this axis. We also illustrate how our measure can be indicative of a conversation's progress, as well as its effectiveness.

Detecting Perceived Emotions in Hurricane Disasters [Website][PDF]
Shrey Desai, Cornelia Caragea, and Junyi Jessy Li 18:00–19:00

Natural disasters (e.g., hurricanes) affect millions of people each year, causing widespread destruction in their wake. People have recently taken to social media websites (e.g., Twitter) to share their sentiments and feelings with the larger community. Consequently, these platforms have become instrumental in understanding and perceiving emotions at scale. In this paper, we introduce HurricaneEmo, an emotion dataset of 15,000 English tweets spanning three hurricanes: Harvey, Irma, and Maria. We present a comprehensive study of fine-grained emotions and propose classification tasks to discriminate between coarse-grained emotion groups. Our best BERT model, even after task-guided pre-training which leverages unlabeled Twitter data, achieves only 68% accuracy (averaged across all groups). HurricaneEmo serves not only as a challenging benchmark for models but also as a valuable resource for analyzing emotions in disaster-centric domains.

Hierarchical Modeling for User Personality Prediction: The Role of Message-Level Attention [Website][PDF]
Veronica Lynn, Niranjan Balasubramanian, and H. Andrew Schwartz 18:00–19:00

Not all documents are equally important. Language processing is increasingly finding use as a supplement for questionnaires to assess psychological attributes of consenting individuals, but most approaches neglect to consider whether all documents of an individual are equally informative. In this paper, we present a novel model that uses message-level attention to learn the relative weight of users' social media posts for assessing their five factor personality traits. We demonstrate that models with message-level attention outperform those with word-level attention, and ultimately yield state-of-the-art accuracies for all five traits by using both word and message attention in combination with past approaches (an average increase in Pearson r of 2.5%). In addition, examination of the high-signal posts identified by our model provides insight into the relationship between language and personality, helping to inform future work.

Measuring Forecasting Skill from Text [Website][PDF]
Shi Zong, Alan Ritter, and Eduard Hovy 18:00–19:00

People vary in their ability to make accurate predictions about the future. Prior studies have shown that some individuals can predict the outcome of future events with consistently better accuracy. This leads to a natural question: what makes some forecasters better than others? In this paper we explore connections between the language people use to describe their predictions and their forecasting skill. Datasets from two different forecasting domains are explored: (1) geopolitical forecasts from Good Judgment Open, an online prediction forum and (2) a corpus of company earnings forecasts made by financial analysts. We present a number of linguistic metrics which are computed over text associated with people's predictions about the future including: uncertainty, readability, and emotion. By

studying linguistic factors associated with predictions, we are able to shed some light on the approach taken by skilled forecasters. Furthermore, we demonstrate that it is possible to accurately predict forecasting skill using a model that is based solely on language. This could potentially be useful for identifying accurate predictions or potentially skilled forecasters earlier.

Text and Causal Inference: A Review of Using Text to Remove Confounding from Causal Estimates [Website][PDF]
Katherine Keith, David Jensen, and Brendan O'Connor 18:00–19:00

Many applications of computational social science aim to infer causal conclusions from non-experimental data. Such observational data often contains confounders, variables that influence both potential causes and potential effects. Unmeasured or latent confounders can bias causal estimates, and this has motivated interest in measuring potential confounders from observed text. For example, an individual's entire history of social media posts or the content of a news article could provide a rich measurement of multiple confounders.Yet, methods and applications for this problem are scattered across different communities and evaluation practices are inconsistent.This review is the first to gather and categorize these examples and provide a guide to data-processing and evaluation decisions. Despite increased attention on adjusting for confounding using text, there are still many open problems, which we highlight in this paper.

Text-Based Ideal Points [Website][PDF]
Keyon Vafa, Suresh Naidu, and David Blei 18:00–19:00

Ideal point models analyze lawmakers' votes to quantify their political positions, or ideal points. But votes are not the only way to express a political position. Lawmakers also give speeches, release press statements, and post tweets. In this paper, we introduce the text-based ideal point model (TBIP), an unsupervised probabilistic topic model that analyzes texts to quantify the political positions of its authors. We demonstrate the TBIP with two types of politicized text data: U.S. Senate speeches and senator tweets. Though the model does not analyze their votes or political affiliations, the TBIP separates lawmakers by party, learns interpretable politicized topics, and infers ideal points close to the classical vote-based ideal points. One benefit of analyzing texts, as opposed to votes, is that the TBIP can estimate ideal points of anyone who authors political texts, including non-voting actors. To this end, we use it to study tweets from the 2020 Democratic presidential candidates. Using only the texts of their tweets, it identifies them along an interpretable progressive-to-moderate spectrum.

Understanding the Language of Political Agreement and Disagreement in Legislative Texts [Website][PDF]
Maryam Davoodi, Eric Waltenburg, and Dan Goldwasser 18:00–19:00

While national politics often receive the spotlight, the overwhelming majority of legislation proposed, discussed, and enacted is done at the state level. Despite this fact, there is little awareness of the dynamics that lead to adopting these policies. In this paper, we take the first step towards a better understanding of these processes and the underlying dynamics that shape them, using data-driven methods. We build a new large-scale dataset, from multiple data sources, connecting state bills and legislator information, geographical information about their districts, and donations and donors' information. We suggest a novel task, predicting the legislative body's vote breakdown for a given bill, according to different criteria of interest, such as gender, rural-urban and ideological splits. Finally, we suggest a shared relational embedding model, representing the interactions between the text of the bill and the legislative context in which it is presented. Our experiments show that providing this context helps improve the prediction over strong text-based models.

Would you Rather? A New Benchmark for Learning Machine Alignment with Cultural Values and Social Preferences [Website][PDF]
Yi Tay, Donovan Ong, Jie Fu, Alvin Chan, Nancy Chen, Anh Tuan Luu, and Chris Pal 18:00–19:00

Understanding human preferences, along with cultural and social nuances, lives at the heart of natural language understanding. Concretely, we present a new task and corpus for learning alignments between machine and human preferences. Our newly introduced problem is concerned with predicting the preferable options from two sentences describing scenarios that may involve social, cultural, ethical, or moral situations. Our problem is framed as a natural language inference task with crowd-sourced preference votes by human players, obtained from a gamified voting platform. Along with the release of a new dataset of 200K data points, we benchmark several state-of-the-art neural models, along with BERT and friends on this task. Our experimental results show that current state-of-the-art NLP models still leave much room for improvement.

Session 9B: Discourse and Pragmatics-4

Discourse as a Function of Event: Profiling Discourse Structure in News Articles around the Main Event [Website][PDF]
Prafulla Kumar Choubey, Aaron Lee, Ruihong Huang, and Lu Wang 18:00–19:00

Understanding discourse structures of news articles is vital to effectively contextualize the occurrence of a news event. To enable computational modeling of news structures, we apply an existing theory of functional discourse structure for news articles that revolves around the main event and create a human-annotated corpus of 802 documents spanning over four domains and three media sources. Next, we propose several document-level neural-network models to automatically construct news content structures. Finally, we demonstrate that incorporating system predicted news structures yields new state-of-the-art performance for event coreference resolution. The news documents we annotated are openly available and the annotations are publicly released for future research.

Harnessing the linguistic signal to predict scalar inferences [Website][PDF]
Sebastian Schuster, Yuxing Chen, and Judith Degen 18:00–19:00

Pragmatic inferences often subtly depend on the presence or absence of linguistic features. For example, the presence of a partitive construction (of the) increases the strength of a so-called scalar inference: listeners perceive the inference that Chris did not eat all of the cookies to be stronger after hearing "Chris ate some of the cookies" than after hearing the same utterance without a partitive, "Chris ate some cookies". In this work, we explore to what extent neural network sentence encoders can learn to predict the strength of scalar inferences. We first show that an LSTM-based sentence encoder trained on an English dataset of human inference strength ratings is able to predict ratings with high accuracy ($r = 0.78$). We then probe the model's behavior using manually constructed minimal sentence pairs and corpus data. We first that the model inferred previously established associations between linguistic features and inference strength, suggesting that the model learns to use linguistic features to predict pragmatic inferences.

Implicit Discourse Relation Classification: We Need to Talk about Evaluation [Website][PDF]
Najoung Kim, Song Feng, Chulaka Gunasekara, and Luis Lastras 18:00–19:00

Implicit relation classification on Penn Discourse TreeBank (PDTB) 2.0 is a common benchmark task for evaluating the understanding of discourse relations. However, the lack of consistency in preprocessing and evaluation poses challenges to fair comparison of results in the literature. In this work, we highlight these inconsistencies and propose an improved evaluation protocol. Paired with this protocol, we report strong baseline results from pretrained sentence encoders, which set the new state-of-the-art for PDTB 2.0. Furthermore, this work is the first to explore fine-grained relation classification on PDTB 3.0. We expect our work to serve as a point of comparison for future work, and also as an initiative to discuss models of larger context and possible data augmentations for downstream transferability.

PeTra: A Sparsely Supervised Memory Model for People Tracking [Website][PDF]
Shubham Toshniwal, Allyson Ettinger, Kevin Gimpel, and Karen Livescu 18:00–19:00

We propose PeTra, a memory-augmented neural network designed to track entities in its memory slots. PeTra is trained using sparse annotation from the GAP pronoun resolution dataset and outperforms a prior memory model on the task while using a simpler architecture. We empirically compare key modeling choices, finding that we can simplify several aspects of the design of the memory module while retaining strong performance. To measure the people tracking capability of memory models, we (a) propose a new diagnostic evaluation based on counting the number of unique entities in text, and (b) conduct a small scale human evaluation to compare evidence of people tracking in the memory logs of PeTra relative to a previous approach. PeTra is highly effective in both evaluations, demonstrating its ability to track people in its memory despite being trained with limited annotation.

ZPR2: Joint Zero Pronoun Recovery and Resolution using Multi-Task Learning and BERT [Website][PDF]
Linfeng Song, Kun Xu, Yue Zhang, Jianshu Chen, and Dong Yu 18:00–19:00

Zero pronoun recovery and resolution aim at recovering the dropped pronoun and pointing out its anaphoric mentions, respectively. We propose to better explore their interaction by solving both tasks together, while the previous work treats them separately. For zero pronoun resolution, we study this task in a more realistic setting, where no parsing trees or only automatic trees are available, while most previous work assumes gold trees. Experiments on two benchmarks show that joint modeling significantly outperforms our baseline that already beats the previous state of the arts.

Session 9B: Ethics and NLP-4

Contextualizing Hate Speech Classifiers with Post-hoc Explanation [Website] [PDF]
Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren 18:00–19:00

Hate speech classifiers trained on imbalanced datasets struggle to determine if group identifiers like "gay" or "black" are used in offensive or prejudiced ways. Such biases manifest in false positives when these identifiers are present, due to models' inability to learn the contexts which constitute a hateful usage of identifiers. We extract post-hoc explanations from fine-tuned BERT classifiers to detect bias towards identity terms. Then, we propose a novel regularization technique based on these explanations that encourages models to learn from the context of group identifiers in addition to the identifiers themselves. Our approach improved over baselines in limiting false positives on out-of-domain data while maintaining and in cases improving in-domain performance.

Double-Hard Debias: Tailoring Word Embeddings for Gender Bias Mitigation [Website] [PDF]
Tianlu Wang, Xi Victoria Lin, Nazneen Fatema Rajani, Bryan McCann, Vicente Ordonez, and Caiming Xiong 18:00–19:00

Word embeddings derived from human-generated corpora inherit strong gender bias which can be further amplified by downstream models. Some commonly adopted debiasing approaches, including the seminal Hard Debias algorithm, apply post-processing procedures that project pre-trained word embeddings into a subspace orthogonal to an inferred gender subspace. We discover that semantic-agnostic corpus regularities such as word frequency captured by the word embeddings negatively impact the performance of these algorithms. We propose a simple but effective technique, Double Hard Debias, which purifies the word embeddings against such corpus regularities prior to inferring and removing the gender subspace. Experiments on three bias mitigation benchmarks show that our approach preserves the distributional semantics of the pre-trained word embeddings while reducing gender bias to a significantly larger degree than prior approaches.

Language (Technology) is Power: A Critical Survey of "Bias" in NLP [Website] [PDF]
Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach 18:00–19:00

We survey 146 papers analyzing "bias" in NLP systems, finding that their motivations are often vague, inconsistent, and lacking in normative reasoning, despite the fact that analyzing "bias" is an inherently normative process. We further find that these papers' proposed quantitative techniques for measuring or mitigating "bias" are poorly matched to their motivations and do not engage with the relevant literature outside of NLP. Based on these findings, we describe the beginnings of a path forward by proposing three recommendations that should guide work analyzing "bias" in NLP systems. These recommendations rest on a greater recognition of the relationships between language and social hierarchies, encouraging researchers and practitioners to articulate their conceptualizations of "bias"—i.e., what kinds of system behaviors are harmful, in what ways, to whom, and why, as well as the normative reasoning underlying these statements—and to center work around the lived experiences of members of communities affected by NLP systems, while interrogating and reimagining the power relations between technologists and such communities.

Social Bias Frames: Reasoning about Social and Power Implications of Language [Website] [PDF]
Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi 18:00–19:00

Warning: this paper contains content that may be offensive or upsetting. Language has the power to reinforce stereotypes and project social biases onto others. At the core of the challenge is that it is rarely what is stated explicitly, but rather the implied meanings, that frame people's judgments about others. For example, given a statement that "we shouldn't lower our standards to hire more women," most listeners will infer the implicature intended by the speaker - that "women (candidates) are less qualified." Most semantic formalisms, to date, do not capture such pragmatic implications in which people express social biases and power differentials in language. We introduce Social Bias Frames, a new conceptual formalism that aims to model the pragmatic frames in which people project social biases and stereotypes onto others. In addition, we introduce the Social Bias Inference Corpus to support large-scale modelling and evaluation with 150k structured annotations of social media posts, covering over 34k implications about a thousand demographic groups. We then establish baseline approaches that learn to recover Social Bias Frames from unstructured text. We find that while state-of-the-art neural models are effective at high-level categorization of whether a given statement projects unwanted social bias (80% F1), they are not effective at spelling out more detailed explanations in terms of Social Bias Frames. Our study motivates future work that combines structured pragmatic inference with commonsense reasoning on social implications.

Social Biases in NLP Models as Barriers for Persons with Disabilities [Website] [PDF]
Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl 18:00–19:00

Building equitable and inclusive NLP technologies demands consideration of whether and how social attitudes are represented in ML models. In particular, representations encoded in models often inadvertently perpetuate undesirable social biases from the data on which they are trained. In this paper, we present evidence of such undesirable biases towards mentions of disability in two different English language models: toxicity prediction and sentiment analysis. Next, we demonstrate that the neural embeddings that are the critical first step in most NLP pipelines similarly contain undesirable biases towards mentions of disability. We end by highlighting topical biases in the discourse about disability which may contribute to the observed model biases; for instance, gun violence, homelessness, and drug addiction are over-represented in texts discussing mental illness.

Towards Debiasing Sentence Representations [Website] [PDF]
Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency 18:00–19:00

As natural language processing methods are increasingly deployed in real-world scenarios such as healthcare, legal systems, and social science, it becomes necessary to recognize the role they potentially play in shaping social biases and stereotypes. Previous work has revealed the presence of social biases in widely used word embeddings involving gender, race, religion, and other social constructs. While some methods were proposed to debias these word-level embeddings, there is a need to perform debiasing at the sentence-level given the recent shift towards new contextualized sentence representations such as ELMo and BERT. In this paper, we investigate the presence of social biases in sentence-level representations and propose a new method, Sent-Debias, to reduce these biases. We show that Sent-Debias is effective in removing biases, and at the same time, preserves performance on sentence-level downstream tasks such as sentiment analysis, linguistic acceptability, and natural language understanding. We hope that our work will inspire future research on characterizing and removing social biases from widely adopted sentence representations for fairer NLP.

Session 9B: Interpretability and Analysis of Models for NLP-6

A Re-evaluation of Knowledge Graph Completion Methods [Website][PDF]
Zhiqing Sun, Shikhar Vashishth, Soumya Sanyal, Partha Talukdar, and Yiming Yang 18:00–19:00

Knowledge Graph Completion (KGC) aims at automatically predicting missing links for large-scale knowledge graphs. A vast number of state-of-the-art KGC techniques have got published at top conferences in several research fields, including data mining, machine learning, and natural language processing. However, we notice that several recent papers report very high performance, which largely outperforms previous state-of-the-art methods. In this paper, we find that this can be attributed to the inappropriate evaluation protocol used by them and propose a simple evaluation protocol to address this problem. The proposed protocol is robust to handle bias in the model, which can substantially affect the final results. We conduct extensive experiments and report performance of several existing methods using our protocol. The reproducible code has been made publicly available.

Cross-Linguistic Syntactic Evaluation of Word Prediction Models [Website][PDF]
Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen 18:00–19:00

A range of studies have concluded that neural word prediction models can distinguish grammatical from ungrammatical sentences with high accuracy. However, these studies are based primarily on monolingual evidence from English. To investigate how these models' ability to learn syntax varies by language, we introduce CLAMS (Cross-Linguistic Assessment of Models on Syntax), a syntactic evaluation suite for monolingual and multilingual models. CLAMS includes subject-verb agreement challenge sets for English, French, German, Hebrew and Russian, generated from grammars we develop. We use CLAMS to evaluate LSTM language models as well as monolingual and multilingual BERT. Across languages, monolingual LSTMs achieved high accuracy on dependencies without attractors, and generally poor accuracy on agreement across object relative clauses. On other constructions, agreement accuracy was generally higher in languages with richer morphology. Multilingual models generally underperformed monolingual models. Multilingual BERT showed high syntactic accuracy on English, but noticeable deficiencies in other languages.

[TACL] Does Syntax Need to Grow on Trees? Sources of Hierarchical Inductive Bias in Sequence-to-Sequence Networks [Website][PDF]
R. Thomas McCoy, Robert Frank, and Tal Linzen 18:00–19:00

Learners that are exposed to the same training data might generalize differently due to differing inductive biases. In neural network models, inductive biases could in theory arise from any aspect of the model architecture. We investigate which architectural factors affect the generalization behavior of neural sequence-to-sequence models trained on two syntactic tasks, English question formation and English tense reinflection. For both tasks, the training set is consistent with a generalization based on hierarchical structure and a generalization based on linear order. All architectural factors that we investigated qualitatively affected how models generalized, including factors with no clear connection to hierarchical structure. For example, LSTMs and GRUs displayed qualitatively different inductive biases. However, the only factor that consistently contributed a hierarchical bias across tasks was the use of a tree-structured model rather than a model with sequential recurrence, suggesting that human-like syntactic generalization requires architectural syntactic structure.

Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior? [Website][PDF]
Peter Hase and Mohit Bansal 18:00–19:00

Algorithmic approaches to interpreting machine learning models have proliferated in recent years. We carry out human subject tests that are the first of their kind to isolate the effect of algorithmic explanations on a key aspect of model interpretability, simulatability, while avoiding important confounding experimental factors. A model is simulatable when a person can predict its behavior on new inputs. Through two kinds of simulation tests involving text and tabular data, we evaluate five explanations methods: (1) LIME, (2) Anchor, (3) Decision Boundary, (4) a Prototype model, and (5) a Composite approach that combines explanations from each method. Clear evidence of method effectiveness is found in very few cases: LIME improves simulatability in tabular classification, and our Prototype method is effective in counterfactual simulation tests. We also collect subjective ratings of explanations, but we do not find that ratings are predictive of how helpful explanations are. Our results provide the first reliable and comprehensive estimates of how explanations influence simulatability across a variety of explanation methods and data domains. We show that (1) we need to be careful about the metrics we use to evaluate explanation methods, and (2) there is significant room for improvement in current methods.

Explaining Black Box Predictions and Unveiling Data Artifacts through Influence Functions [Website][PDF]
Xiaochuang Han, Byron C. Wallace, and Yulia Tsvetkov 18:00–19:00

Modern deep learning models for NLP are notoriously opaque. This has motivated the development of methods for interpreting such models, e.g., via gradient-based saliency maps or the visualization of attention weights. Such approaches aim to provide explanations for a particular model prediction by highlighting important words in the corresponding input text. While this might be useful for tasks where decisions are explicitly influenced by individual tokens in the input, we suspect that such highlighting is not suitable for tasks where model decisions should be driven by more complex reasoning. In this work, we investigate the use of influence functions for NLP, providing an alternative approach to interpreting neural text classifiers. Influence functions explain the decisions of a model by identifying influential training examples. Despite the promise of this approach, influence functions have not yet been extensively evaluated in the context of NLP, a gap addressed by this work. We conduct a comparison between influ-

ence functions and common word-saliency methods on representative tasks. As suspected, we find that influence functions are particularly useful for natural language inference, a task in which 'saliency maps' may not have clear interpretation. Furthermore, we develop a new quantitative measure based on influence functions that can reveal artifacts in training data.

Finding Universal Grammatical Relations in Multilingual BERT [Website][PDF]
Ethan A. Chi, John Hewitt, and Christopher D. Manning 18:00–19:00

Recent work has found evidence that Multilingual BERT (mBERT), a transformer-based multilingual masked language model, is capable of zero-shot cross-lingual transfer, suggesting that some aspects of its representations are shared cross-lingually. To better understand this overlap, we extend recent work on finding syntactic trees in neural networks' internal representations to the multilingual setting. We show that subspaces of mBERT representations recover syntactic tree distances in languages other than English, and that these subspaces are approximately shared across languages. Motivated by these results, we present an unsupervised analysis method that provides evidence mBERT learns representations of syntactic dependency labels, in the form of clusters which largely agree with the Universal Dependencies taxonomy. This evidence suggests that even without explicit supervision, multilingual masked language models learn certain linguistic universals.

Generating Hierarchical Explanations on Text Classification via Feature Interaction Detection [Website][PDF]
Hanjie Chen, Guangtao Zheng, and Yangfeng Ji 18:00–19:00

Generating explanations for neural networks has become crucial for their applications in real-world with respect to reliability and trustworthiness. In natural language processing, existing methods usually provide important features which are words or phrases selected from an input text as an explanation, but ignore the interactions between them. It poses challenges for humans to interpret an explanation and connect it to model prediction. In this work, we build hierarchical explanations by detecting feature interactions. Such explanations visualize how words and phrases are combined at different levels of the hierarchy, which can help users understand the decision-making of black-box models. The proposed method is evaluated with three neural text classifiers (LSTM, CNN, and BERT) on two benchmark datasets, via both automatic and human evaluations. Experiments show the effectiveness of the proposed method in providing explanations that are both faithful to models and interpretable to humans.

Obtaining Faithful Interpretations from Compositional Neural Networks [Website][PDF]
Sanjay Subramanian, Ben Bogin, Nitish Gupta, Tomer Wolfson, Sameer Singh, Jonathan Berant, and Matt Gardner 18:00–19:00

Neural module networks (NMNs) are a popular approach for modeling compositionality: they achieve high accuracy when applied to problems in language and vision, while reflecting the compositional structure of the problem in the network architecture. However, prior work implicitly assumed that the structure of the network modules, describing the abstract reasoning process, provides a faithful explanation of the model's reasoning; that is, that all modules perform their intended behaviour. In this work, we propose and conduct a systematic evaluation of the intermediate outputs of NMNs on NLVR2 and DROP, two datasets which require composing multiple reasoning steps. We find that the intermediate outputs differ from the expected output, illustrating that the network structure does not provide a faithful explanation of model behaviour. To remedy that, we train the model with auxiliary supervision and propose particular choices for module architecture that yield much better faithfulness, at a minimal cost to accuracy.

On the Cross-lingual Transferability of Monolingual Representations [Website][PDF]
Mikel Artetxe, Sebastian Ruder, and Dani Yogatama 18:00–19:00

State-of-the-art unsupervised multilingual models (e.g., multilingual BERT) have been shown to generalize in a zero-shot cross-lingual setting. This generalization ability has been attributed to the use of a shared subword vocabulary and joint training across multiple languages giving rise to deep multilingual abstractions. We evaluate this hypothesis by designing an alternative approach that transfers a monolingual model to new languages at the lexical level. More concretely, we first train a transformer-based masked language model on one language, and transfer it to a new language by learning a new embedding matrix with the same masked language modeling objective, freezing parameters of all other layers. This approach does not rely on a shared vocabulary or joint training. However, we show that it is competitive with multilingual BERT on standard cross-lingual classification benchmarks and on a new Cross-lingual Question Answering Dataset (XQuAD). Our results contradict common beliefs of the basis of the generalization ability of multilingual models and suggest that deep monolingual models learn some abstractions that generalize across languages. We also release XQuAD as a more comprehensive cross-lingual benchmark, which comprises 240 paragraphs and 1190 question-answer pairs from SQuAD v1.1 translated into ten languages by professional translators.

Rationalizing Text Matching: Learning Sparse Alignments via Optimal Transport [Website][PDF]
Kyle Swanson, Lili Yu, and Tao Lei 18:00–19:00

Selecting input features of top relevance has become a popular method for building self-explaining models. In this work, we extend this selective rationalization approach to text matching, where the goal is to jointly select and align text pieces, such as tokens or sentences, as a justification for the downstream prediction. Our approach employs optimal transport (OT) to find a minimal cost alignment between the inputs. However, directly applying OT often produces dense and therefore uninterpretable alignments. To overcome this limitation, we introduce novel constrained variants of the OT problem that result in highly sparse alignments with controllable sparsity. Our model is end-to-end differentiable using the Sinkhorn algorithm for OT and can be trained without any alignment annotations. We evaluate our model on the StackExchange, MultiNews, e-SNLI, and MultiRC datasets. Our model achieves very sparse rationale selections with high fidelity while preserving prediction accuracy compared to strong attention baseline models.

Session 9B: Question Answering-6

Benefits of Intermediate Annotations in Reading Comprehension [Website][PDF]
Dheeru Dua, Sameer Singh, and Matt Gardner 18:00–19:00

Complex compositional reading comprehension datasets require performing latent sequential decisions that are learned via supervision from the final answer. A large combinatorial space of possible decision paths that result in the same answer, compounded by the lack of intermediate supervision to help choose the right path, makes the learning particularly hard for this task. In this work, we study the benefits of collecting intermediate reasoning supervision along with the answer during data collection. We find that these intermediate annotations can provide two-fold benefits. First, we observe that for any collection budget, spending a fraction of it on intermediate annotations results in improved model performance, for two complex compositional datasets: DROP and Quoref. Second, these annotations encourage the model to learn the correct latent reasoning steps, helping combat some of the biases introduced during the data collection process.

[TACL] Break It Down: A Question Understanding Benchmark [Website][PDF]
Tomer Wolfson, Mor Geva, Ankit Gupta, Yoav Goldberg, Matt Gardner, Daniel Deutch, and Jonathan Berant 18:00–19:00

Understanding natural language questions entails the ability to break down a question into the requisite steps for computing its answer. In this work, we introduce a Question Decomposition Meaning Representation (QDMR) for questions. QDMR constitutes the ordered list of steps, expressed through natural language, that are necessary for answering a question. We develop a crowdsourcing pipeline, showing that quality QDMRs can be annotated at scale, and release the Break dataset, containing over 83K pairs of questions and their QDMRs.

Crossing Variational Autoencoders for Answer Retrieval [Website][PDF]
Wenhao Yu, Lingfei Wu, Qingkai Zeng, Shu Tao, Yu Deng, and Meng Jiang 18:00–19:00

Answer retrieval is to find the most aligned answer from a large set of candidates given a question. Learning vector representations of questions/answers is the key factor. Question-answer alignment and question/answer semantics are two important signals for learning the representations. Existing methods learned semantic representations with dual encoders or dual variational auto-encoders. The semantic information was learned from language models or question-to-question (answer-to-answer) generative processes. However, the alignment and semantics were too separate to capture the aligned semantics between question and answer. In this work, we propose to cross variational auto-encoders by generating questions with aligned answers and generating answers with aligned questions. Experiments show that our method outperforms the state-of-the-art answer retrieval method on SQuAD.

Improving Multi-hop Question Answering over Knowledge Graphs using Knowledge Base Embeddings [Website][PDF]
Apoorv Saxena, Aditay Tripathi, and Partha Talukdar 18:00–19:00

Knowledge Graphs (KG) are multi-relational graphs consisting of entities as nodes and relations among them as typed edges. Goal of the Question Answering over KG (KGQA) task is to answer natural language queries posed over the KG. Multi-hop KGQA requires reasoning over multiple edges of the KG to arrive at the right answer. KGs are often incomplete with many missing links, posing additional challenges for KGQA, especially for multi-hop KGQA. Recent research on multi-hop KGQA has attempted to handle KG sparsity using relevant external text, which isn't always readily available. In a separate line of research, KG embedding methods have been proposed to reduce KG sparsity by performing missing link prediction. Such KG embedding methods, even though highly relevant, have not been explored for multi-hop KGQA so far. We fill this gap in this paper and propose EmbedKGQA. EmbedKGQA is particularly effective in performing multi-hop KGQA over sparse KGs. EmbedKGQA also relaxes the requirement of answer selection from a pre-specified neighborhood, a sub-optimal constraint enforced by previous multi-hop KGQA methods. Through extensive experiments on multiple benchmark datasets, we demonstrate EmbedKGQA's effectiveness over other state-of-the-art baselines.

[TACL] Investigating Prior Knowledge for Challenging Chinese Machine Reading Comprehension [Website][PDF]
Kai Sun, Dian Yu, Dong Yu, and Claire Cardie 18:00–19:00

Machine reading comprehension tasks require a machine reader to answer questions relevant to the given document. In this paper, we present the first free-form multiple-Choice Chinese machine reading Comprehension dataset (C^3), containing 13,369 documents (dialogues or more formally written mixed-genre texts) and their associated 19,577 multiple-choice free-form questions collected from Chinese-as-a-second-language examinations.

Logic-Guided Data Augmentation and Regularization for Consistent Question Answering [Website][PDF]
Akari Asai and Hannaneh Hajishirzi 18:00–19:00

Many natural language questions require qualitative, quantitative or logical comparisons between two entities or events. This paper addresses the problem of improving the accuracy and consistency of responses to comparison questions by integrating logic rules and neural models. Our method leverages logical and linguistic knowledge to augment labeled training data and then uses a consistency-based regularizer to train the model. Improving the global consistency of predictions, our approach achieves large improvements over previous methods in a variety of question answering (QA) tasks, including multiple-choice qualitative reasoning, cause-effect reasoning, and extractive machine reading comprehension. In particular, our method significantly improves the performance of RoBERTa-based

models by 1-5% across datasets. We advance state of the art by around 5-8% on WIQA and QuaRel and reduce consistency violations by 58% on HotpotQA. We further demonstrate that our approach can learn effectively from limited data.

On the Importance of Diversity in Question Generation for QA [Website][PDF]
Md Arafat Sultan, Shubham Chandel, Ramón Fernandez Astudillo, and Vittorio Castelli 18:00–19:00

Automatic question generation (QG) has shown promise as a source of synthetic training data for question answering (QA). In this paper we ask: Is textual diversity in QG beneficial for downstream QA? Using top-p nucleus sampling to derive samples from a transformer-based question generator, we show that diversity-promoting QG indeed provides better QA training than likelihood maximization approaches such as beam search. We also show that standard QG evaluation metrics such as BLEU, ROUGE and METEOR are inversely correlated with diversity, and propose a diversity-aware intrinsic measure of overall QG quality that correlates well with extrinsic evaluation on QA.

Probabilistic Assumptions Matter: Improved Models for Distantly-Supervised Document-Level Question Answering [Website][PDF]
Hao Cheng, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova 18:00–19:00

We address the problem of extractive question answering using document-level distant super-vision, pairing questions and relevant documents with answer strings. We compare previously used probability space and distant supervision assumptions (assumptions on the correspondence between the weak answer string labels and possible answer mention spans). We show that these assumptions interact, and that different configurations provide complementary benefits. We demonstrate that a multi-objective model can efficiently combine the advantages of multiple assumptions and outperform the best individual formulation. Our approach outperforms previous state-of-the-art models by 4.3 points in F1 on TriviaQA-Wiki and 1.7 points in Rouge-L on NarrativeQA summaries.

SCDE: Sentence Cloze Dataset with High Quality Distractors From Examinations [Website][PDF]
Xiang Kong, Varun Gangal, and Eduard Hovy 18:00–19:00

We introduce SCDE, a dataset to evaluate the performance of computational models through sentence prediction. SCDE is a human created sentence cloze dataset, collected from public school English examinations. Our task requires a model to fill up multiple blanks in a passage from a shared candidate set with distractors designed by English teachers. Experimental results demonstrate that this task requires the use of non-local, discourse-level context beyond the immediate sentence neighborhood. The blanks require joint solving and significantly impair each other's context. Furthermore, through ablations, we show that the distractors are of high quality and make the task more challenging. Our experiments show that there is a significant performance gap between advanced models (72%) and humans (87%), encouraging future models to bridge this gap.

Selective Question Answering under Domain Shift [Website][PDF]
Amita Kamath, Robin Jia, and Percy Liang 18:00–19:00

To avoid giving wrong answers, question answering (QA) models need to know when to abstain from answering. Moreover, users often ask questions that diverge from the model's training data, making errors more likely and thus abstention more critical. In this work, we propose the setting of selective question answering under domain shift, in which a QA model is tested on a mixture of in-domain and out-of-domain data, and must answer (i.e., not abstain on) as many questions as possible while maintaining high accuracy. Abstention policies based solely on the model's softmax probabilities fare poorly, since models are overconfident on out-of-domain inputs. Instead, we train a calibrator to identify inputs on which the QA model errs, and abstain when it predicts an error is likely. Crucially, the calibrator benefits from observing the model's behavior on out-of-domain data, even if from a different domain than the test data. We combine this method with a SQuAD-trained QA model and evaluate on mixtures of SQuAD and five other QA datasets. Our method answers 56% of questions while maintaining 80% accuracy; in contrast, directly using the model's probabilities only answers 48% at 80% accuracy.

Template-Based Question Generation from Retrieved Sentences for Improved Unsupervised Question Answering [Website][PDF]
Alexander Fabbri, Patrick Ng, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang 18:00–19:00

Question Answering (QA) is in increasing demand as the amount of information available online and the desire for quick access to this content grows. A common approach to QA has been to fine-tune a pretrained language model on a task-specific labeled dataset. This paradigm, however, relies on scarce, and costly to obtain, large-scale human-labeled data. We propose an unsupervised approach to training QA models with generated pseudo-training data. We show that generating questions for QA training by applying a simple template on a related, retrieved sentence rather than the original context sentence improves downstream QA performance by allowing the model to learn more complex context-question relationships. Training a QA model on this data gives a relative improvement over a previous unsupervised model in F1 score on the SQuAD dataset by about 14%, and 20% when the answer is a named entity, achieving state-of-the-art performance on SQuAD for unsupervised QA.

The Cascade Transformer: an Application for Efficient Answer Sentence Selection [Website][PDF]
Luca Soldaini and Alessandro Moschitti 18:00–19:00

Large transformer-based language models have been shown to be very effective in many classification tasks. However, their computational complexity prevents their use in applications requiring the classification of a large set of candidates. While previous works have investigated approaches to reduce model size, relatively little attention has been paid to techniques to improve batch throughput during inference. In this paper, we introduce the Cascade Transformer, a simple yet effective technique to adapt transformer-based models into a cascade of rankers. Each ranker is used to prune a subset of candidates in a batch, thus dramatically increasing throughput at inference time. Partial

encodings from the transformer model are shared among rerankers, providing further speed-up. When compared to a state-of-the-art transformer model, our approach reduces computation by 37% with almost no impact on accuracy, as measured on two English Question Answering datasets.

Transformers to Learn Hierarchical Contexts in Multiparty Dialogue for Span-based Question Answering [Website][PDF]
Changmao Li and Jinho D. Choi 18:00–19:00

We introduce a novel approach to transformers that learns hierarchical representations in multiparty dialogue. First, three language modeling tasks are used to pre-train the transformers, token- and utterance-level language modeling and utterance order prediction, that learn both token and utterance embeddings for better understanding in dialogue contexts. Then, multi-task learning between the utterance prediction and the token span prediction is applied to fine-tune for span-based question answering (QA). Our approach is evaluated on the FriendsQA dataset and shows improvements of 3.8% and 1.4% over the two state-of-the-art transformer models, BERT and RoBERTa, respectively.

[TACL] TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages [Website][PDF]
Jonathan H Clark, Jennimaria Palomaki, Vitaly Nikolaev, Eunsol Choi, Dan Garrette, Michael Collins, and Tom Kwiatkowski 18:00–19:00

Confidently making progress on multilingual modeling requires challenging, trustworthy evaluations. We present TyDi QA, a question answering dataset covering 11 typologically diverse languages with 141K question-answer pairs. The languages of TyDi QA are diverse with regard to their typology — the set of linguistic features that each language expresses — such that we expect models performing well on this set to generalize across a large number of the languages in the world. We present a quantitative analysis of the data quality and example-level qualitative linguistic analyses of observed language phenomena that would not be found in English-only corpora. To provide a realistic information-seeking task and avoid priming effects, questions are written by people who want to know the answer, but don't know the answer yet, and the data is collected directly in each language without the use of translation. We provide initial quality measurements with a baseline model, suggesting a significant room for future work on this data.

Session 9B: Resources and Evaluation-10

A Corpus for Large-Scale Phonetic Typology [Website][PDF]
Elizabeth Salesky, Eleanor Chodroff, Tiago Pimentel, Matthew Wiesner, Ryan Cotterell, Alan W Black, and Jason Eisner 18:00–19:00

A major hurdle in data-driven research on typology is having sufficient data in many languages to draw meaningful conclusions. We present VoxClamantis v1.0, the first large-scale corpus for phonetic typology, with aligned segments and estimated phoneme-level labels in 690 readings spanning 635 languages, along with acoustic-phonetic measures of vowels and sibilants. Access to such data can greatly facilitate investigation of phonetic typology at a large scale and across many languages. However, it is non-trivial and computationally intensive to obtain such alignments for hundreds of languages, many of which have few to no resources presently available. We describe the methodology to create our corpus, discuss caveats with current methods and their impact on the utility of this data, and illustrate possible research directions through a series of case studies on the 48 highest-quality readings. Our corpus and scripts are publicly available for non-commercial use at https://voxclamantisproject.github.io.

An Effectiveness Metric for Ordinal Classification: Formal Properties and Experimental Results [Website][PDF]
Enrique Amigo, Julio Gonzalo, Stefano Mizzaro, and Jorge Carrillo-de-Albornoz 18:00–19:00

In Ordinal Classification tasks, items have to be assigned to classes that have a relative ordering, such as "positive", "neutral", "negative" in sentiment analysis. Remarkably, the most popular evaluation metrics for ordinal classification tasks either ignore relevant information (for instance, precision/recall on each of the classes ignores their relative ordering) or assume additional information (for instance, Mean Average Error assumes absolute distances between classes). In this paper we propose a new metric for Ordinal Classification, Closeness Evaluation Measure, that is rooted on Measurement Theory and Information Theory. Our theoretical analysis and experimental results over both synthetic data and data from NLP shared tasks indicate that the proposed metric captures quality aspects from different traditional tasks simultaneously. In addition, it generalizes some popular classification (nominal scale) and error minimization (interval scale) metrics, depending on the measurement scale in which it is instantiated.

Not All Claims are Created Equal: Choosing the Right Statistical Approach to Assess Hypotheses [Website][PDF]
Erfan Sadeqi Azer, Daniel Khashabi, Ashish Sabharwal, and Dan Roth 18:00–19:00

Empirical research in Natural Language Processing (NLP) has adopted a narrow set of principles for assessing hypotheses, relying mainly on p-value computation, which suffers from several known issues. While alternative proposals have been well-debated and adopted in other fields, they remain rarely discussed or used within the NLP community. We address this gap by contrasting various hypothesis assessment techniques, especially those not commonly used in the field (such as evaluations based on Bayesian inference). Since these statistical techniques differ in the hypotheses they can support, we argue that practitioners should first decide their target hypothesis before choosing an assessment method. This is crucial because common fallacies, misconceptions, and misinterpretation surrounding hypothesis assessment methods often stem from a discrepancy between what one would like to claim versus what the method used actually assesses. Our survey reveals that these issues are omnipresent in the NLP research community. As a step forward, we provide best practices and guidelines tailored to NLP research, as well as an easy-to-use package for Bayesian assessment of hypotheses, complementing existing tools.

STARC: Structured Annotations for Reading Comprehension [Website][PDF]
Yevgeni Berzak, Jonathan Malmaud, and Roger Levy 18:00–19:00

We present STARC (Structured Annotations for Reading Comprehension), a new annotation framework for assessing reading comprehension with multiple choice questions. Our framework introduces a principled structure for the answer choices and ties them to textual span annotations. The framework is implemented in OneStopQA, a new high-quality dataset for evaluation and analysis of reading comprehension in English. We use this dataset to demonstrate that STARC can be leveraged for a key new application for the development of SAT-like reading comprehension materials: automatic annotation quality probing via span ablation experiments. We further show that it enables in-depth analyses and comparisons between machine and human reading comprehension behavior, including error distributions and guessing ability. Our experiments also reveal that the standard multiple choice dataset in NLP, RACE, is limited in its ability to measure reading comprehension. 47% of its questions can be guessed by machines without accessing the passage, and 18% are unanimously judged by humans as not having a unique correct answer. OneStopQA provides an alternative test set for reading comprehension which alleviates these shortcomings and has a substantially higher human ceiling performance.

WinoWhy: A Deep Diagnosis of Essential Commonsense Knowledge for Answering Winograd Schema Challenge [Website][PDF]
Hongming Zhang, Xinran Zhao, and Yangqiu Song 18:00–19:00

In this paper, we present the first comprehensive categorization of essential commonsense knowledge for answering the Winograd Schema Challenge (WSC). For each of the questions, we invite annotators to first provide reasons for making correct decisions and then categorize them into six major knowledge categories. By doing so, we better understand the limitation of existing methods (i.e., what kind of knowledge cannot be effectively represented or inferred with existing methods) and shed some light on the commonsense knowledge that we need to acquire in the future for better commonsense reasoning. Moreover, to investigate whether current WSC models can understand the commonsense or they simply solve the WSC questions based on the statistical bias of the dataset, we leverage the

collected reasons to develop a new task called WinoWhy, which requires models to distinguish plausible reasons from very similar but wrong reasons for all WSC questions. Experimental results prove that even though pre-trained language representation models have achieved promising progress on the original WSC dataset, they are still struggling at WinoWhy. Further experiments show that even though supervised models can achieve better performance, the performance of these models can be sensitive to the dataset distribution. WinoWhy and all codes are available at: https://github.com/HKUST-KnowComp/WinoWhy.

Session 9B: Sentiment Analysis, Stylistic Analysis, and Argument Mining-7

Agreement Prediction of Arguments in Cyber Argumentation for Detecting Stance Polarity and Intensity [Website][PDF]
Joseph Sirrianni, Xiaoqing Liu, and Douglas Adams 18:00–19:00

In online debates, users express different levels of agreement/disagreement with one another's arguments and ideas. Often levels of agreement/disagreement are implicit in the text, and must be predicted to analyze collective opinions. Existing stance detection methods predict the polarity of a post's stance toward a topic or post, but don't consider the stance's degree of intensity. We introduce a new research problem, stance polarity and intensity prediction in response relationships between posts. This problem is challenging because differences in stance intensity are often subtle and require nuanced language understanding. Cyber argumentation research has shown that incorporating both stance polarity and intensity data in online debates leads to better discussion analysis. We explore five different learning models: Ridge-M regression, Ridge-S regression, SVR-RF-R, pkudblab-PIP, and T-PAN-PIP for predicting stance polarity and intensity in argumentation. These models are evaluated using a new dataset for stance polarity and intensity prediction collected using a cyber argumentation platform. The SVR-RF-R model performs best for prediction of stance polarity with an accuracy of 70.43% and intensity with RMSE of 0.596. This work is the first to train models for predicting a post's stance polarity and intensity in one combined value in cyber argumentation with reasonably good accuracy.

Cross-Lingual Unsupervised Sentiment Classification with Multi-View Transfer Learning [Website][PDF]
Hongliang Fei and Ping Li 18:00–19:00

Recent neural network models have achieved impressive performance on sentiment classification in English as well as other languages. Their success heavily depends on the availability of a large amount of labeled data or parallel corpus. In this paper, we investigate an extreme scenario of cross-lingual sentiment classification, in which the low-resource language does not have any labels or parallel corpus. We propose an unsupervised cross-lingual sentiment classification model named multi-view encoder-classifier (MVEC) that leverages an unsupervised machine translation (UMT) system and a language discriminator. Unlike previous language model (LM) based fine-tuning approaches that adjust parameters solely based on the classification error on training data, we employ the encoder-decoder framework of a UMT as a regularization component on the shared network parameters. In particular, the cross-lingual encoder of our model learns a shared representation, which is effective for both reconstructing input sentences of two languages and generating more representative views from the input for classification. Extensive experiments on five language pairs verify that our model significantly outperforms other models for 8/11 sentiment classification tasks.

Efficient Pairwise Annotation of Argument Quality [Website][PDF]
Lukas Gienapp, Benno Stein, Matthias Hagen, and Martin Potthast 18:00–19:00

We present an efficient annotation framework for argument quality, a feature difficult to be measured reliably as per previous work. A stochastic transitivity model is combined with an effective sampling strategy to infer high-quality labels with low effort from crowdsourced pairwise judgments. The model's capabilities are showcased by compiling Webis-ArgQuality-20, an argument quality corpus that comprises scores for rhetorical, logical, dialectical, and overall quality inferred from a total of 41,859 pairwise judgments among 1,271 arguments. With up to 93% cost savings, our approach significantly outperforms existing annotation procedures. Furthermore, novel insight into argument quality is provided through statistical analysis, and a new aggregation method to infer overall quality from individual quality dimensions is proposed.

Entity-Aware Dependency-Based Deep Graph Attention Network for Comparative Preference Classification [Website][PDF]
Nianzu Ma, Sahisnu Mazumder, Hao Wang, and Bing Liu 18:00–19:00

This paper studies the task of comparative preference classification (CPC). Given two entities in a sentence, our goal is to classify whether the first (or the second) entity is preferred over the other or no comparison is expressed at all between the two entities. Existing works either do not learn entity-aware representations well and fail to deal with sentences involving multiple entity pairs or use sequential modeling approaches that are unable to capture long-range dependencies between the entities. Some also use traditional machine learning approaches that do not generalize well. This paper proposes a novel Entity-aware Dependency-based Deep Graph Attention Network (ED-GAT) that employs a multi-hop graph attention over a dependency graph sentence representation to leverage both the semantic information from word embeddings and the syntactic information from the dependency graph to solve the problem. Empirical evaluation shows that the proposed model achieves the state-of-the-art performance in comparative preference classification.

GoEmotions: A Dataset of Fine-Grained Emotions [Website][PDF]
Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi 18:00–19:00

Understanding emotion expressed in language has a wide range of applications, from building empathetic chatbots to detecting harmful online behavior. Advancement in this area can be improved using large-scale datasets with a fine-grained typology, adaptable to multiple downstream tasks. We introduce GoEmotions, the largest manually annotated dataset of 58k English Reddit comments, labeled for 27 emotion categories or Neutral. We demonstrate the high quality of the annotations via Principal Preserved Component Analysis. We conduct transfer learning experiments with existing emotion benchmarks to show that our dataset generalizes well to other domains and different emotion

taxonomies. Our BERT-based model achieves an average F1-score of .46 across our proposed taxonomy, leaving much room for improvement.

OpinionDigest: A Simple Framework for Opinion Summarization [Website] [PDF]
Yoshihiko Suhara, Xiaolan Wang, Stefanos Angelidis, and Wang-Chiew Tan 18:00–19:00

We present OpinionDigest, an abstractive opinion summarization framework, which does not rely on gold-standard summaries for training. The framework uses an Aspect-based Sentiment Analysis model to extract opinion phrases from reviews, and trains a Transformer model to reconstruct the original reviews from these extractions. At summarization time, we merge extractions from multiple reviews and select the most popular ones. The selected opinions are used as input to the trained Transformer model, which verbalizes them into an opinion summary. OpinionDigest can also generate customized summaries, tailored to specific user needs, by filtering the selected opinions according to their aspect and/or sentiment. Automatic evaluation on Yelp data shows that our framework outperforms competitive baselines. Human studies on two corpora verify that OpinionDigest produces informative summaries and shows promising customization capabilities.

A Comprehensive Analysis of Preprocessing for Word Representation Learning in Affective Tasks [Website] [PDF]
Nastaran Babanejad, Ameeta Agrawal, Aijun An, and Manos Papagelis 18:00–19:00

Affective tasks such as sentiment analysis, emotion classification, and sarcasm detection have been popular in recent years due to an abundance of user-generated data, accurate computational linguistic models, and a broad range of relevant applications in various domains. At the same time, many studies have highlighted the importance of text preprocessing, as an integral step to any natural language processing prediction model and downstream task. While preprocessing in affective systems is well-studied, preprocessing in word vector-based models applied to affective systems, is not. To address this limitation, we conduct a comprehensive analysis of the role of preprocessing techniques in affective analysis based on word vector models. Our analysis is the first of its kind and provides useful insights of the importance of each preprocessing technique when applied at the training phase, commonly ignored in pretrained word vector models, and/or at the downstream task phase.

Session 9B: Student Research Workshop

Compositional Generalization by Factorizing Alignment and Translation [Website][PDF]
Jacob Russin, Jason Jo, Randall O'Reilly, and Yoshua Bengio 18:00–19:00

Standard methods in deep learning for natural language processing fail to capture the compositional structure of human language that allows for systematic generalization outside of the training distribution. However, human learners readily generalize in this way, e.g. by applying known grammatical rules to novel words. Inspired by work in cognitive science suggesting a functional distinction between systems for syntactic and semantic processing, we implement a modification to an existing approach in neural machine translation, imposing an analogous separation between alignment and translation. The resulting architecture substantially outperforms standard recurrent networks on the SCAN dataset, a compositional generalization task, without any additional supervision. Our work suggests that learning to align and to translate in separate modules may be a useful heuristic for capturing compositional structure.

RPD: A Distance Function Between Word Embeddings [Website][PDF]
Xuhui Zhou, Shujian Huang, and Zaixiang Zheng 18:00–19:00

It is well-understood that different algorithms, training processes, and corpora produce different word embeddings. However, less is known about the relation between different embedding spaces, i.e. how far different sets of embeddings deviate from each other. In this paper, we propose a novel metric called Relative Pairwise Inner Product Distance (RPD) to quantify the distance between different sets of word embeddings. This unitary-invariant metric has a unified scale for comparing different sets of word embeddings. Based on the properties of RPD, we study the relations of word embeddings of different algorithms systematically and investigate the influence of different training processes and corpora. The results shed light on the poorly understood word embeddings and justify RPD as a measure of the distance of embedding space.

#NotAWhore! A Computational Linguistic Perspective of Rape Culture and Victimization on Social Media [Website][PDF]
Ashima Suvarna and Grusha Bhalla 18:00–19:00

The recent surge in online forums and movements supporting sexual assault survivors has led to the emergence of a 'virtual bubble' where survivors can recount their stories. However, this also makes the survivors vulnerable to bullying, trolling and victim blaming. Specifically, victim blaming has been shown to have acute psychological effects on the survivors and further discourage formal reporting of such crimes. Therefore, it is important to devise computationally relevant methods to identify and prevent victim blaming to protect the victims. In our work, we discuss the drastic effects of victim blaming through a short case study and then propose a single step transfer-learning based classification method to identify victim blaming language on Twitter. Finally, we compare the performance of our proposed model against various deep learning and machine learning models on a manually annotated domain-specific dataset.

Research Replication Prediction Using Weakly Supervised Learning [Website]
Tianyi Luo, Xingyu Li, Hainan Wang, and Yang Liu 18:00–19:00

Knowing whether a published research result can be replicated or not is important. Carrying out direct replication of published research incurs high cost. It is therefore desirable to have a machine learning aided automatic prediction of a result's replicability. Such predictions can provide a confidence score for each article which can further provide guidelines for spot-checks. Since we will only have access to a small size of annotated dataset to train a machine predictor, we explore the possibility of using weakly supervised learning approaches to improve the prediction accuracy of research replication using both labelled and unlabelled datasets based on text information of research papers. Our experiments over real-world datasets show that much better prediction performance can be obtained compared to the supervised models utilizing only a small size of labelled dataset.

Inducing Grammar from Long Short-Term Memory Networks by Shapley Decomposition [Website][PDF]
Yuhui Zhang and Allen Nie 18:00–19:00

The principle of compositionality has deep roots in linguistics: the meaning of an expression is determined by its structure and the meanings of its constituents. However, modern neural network models such as long short-term memory network process expressions in a linear fashion and do not seem to incorporate more complex compositional patterns. In this work, we show that we can explicitly induce grammar by tracing the computational process of a long short-term memory network. We show: (i) the multiplicative nature of long short-term memory network allows complex interaction beyond sequential linear combination; (ii) we can generate compositional trees from the network without external linguistic knowledge; (iii) we evaluate the syntactic difference between the generated trees, randomly generated trees and gold reference trees produced by constituency parsers; (iv) we evaluate whether the generated trees contain the rich semantic information.

Demo Session 4C

Time: *18:30–19:15*

NLP Scholar: An Interactive Visual Explorer for Natural Language Processing Literature [Website][PDF]
Saif M. Mohammad

As part of the NLP Scholar project, we created a single unified dataset of NLP papers and their meta-information (including citation numbers), by extracting and aligning information from the ACL Anthology and Google Scholar. In this paper, we describe several interconnected interactive visualizations (dashboards) that present various aspects of the data. Clicking on an item within a visualization or entering query terms in the search boxes filters the data in all visualizations in the dashboard. This allows users to search for papers in the area of their interest, published within specific time periods, published by specified authors, etc. The interactive visualizations presented here, and the associated dataset of papers mapped to citations, have additional uses as well including understanding how the field is growing (both overall and across sub-areas), as well as quantifying the impact of different types of papers on subsequent publications.

Stimulating Creativity with FunLines: A Case Study of Humor Generation in Headlines [Website][PDF]
Nabil Hossain, John Krumm, Tanvir Sajed, and Henry Kautz

Building datasets of creative text, such as humor, is quite challenging. We introduce FunLines, a competitive game where players edit news headlines to make them funny, and where they rate the funniness of headlines edited by others. FunLines makes the humor generation process fun, interactive, collaborative, rewarding and educational, keeping players engaged and providing humor data at a very low cost compared to traditional crowdsourcing approaches. FunLines offers useful performance feedback, assisting players in getting better over time at generating and assessing humor, as our analysis shows. This helps to further increase the quality of the generated dataset. We show the effectiveness of this data by training humor classification models that outperform a previous benchmark, and we release this dataset to the public.

Usnea: An Authorship Tool for Interactive Fiction using Retrieval Based Semantic Parsing [Website][PDF]
Ben Swanson and Boris Smus

The reader of a choose your own adventure novel and the user of a modern virtual assistant have a subtle similarity; both may, through the right lens, be viewed as engaging with a work of Interactive Fiction. This literary form emerged in the 1970s and has grown like a vine along the branch of modern technology, one guided by the advances of the other. In this work we weave together threads from the Interactive Fiction community and neural semantic parsing for dialog systems, defining the data model and necessary algorithms for a novel type of Interactive Fiction and open sourcing its accompanying authoring tool. Specifically, our work integrates retrieval based semantic parsing predicates into the branching story structures well known to the Interactive Fiction community, relaxing the relatively strict lexical options of preexisting systems.

Demo Session 5A

Time: *20:00–20:45*

DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation [Website][PDF]

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan

We present a large, tunable neural conversational response generation model, DIALOGPT (dialogue generative pre-trained transformer). Trained on 147M conversation-like exchanges extracted from Reddit comment chains over a period spanning from 2005 through 2017, DialoGPT extends the Hugging Face PyTorch transformer to attain a performance close to human both in terms of automatic and human evaluation in single-turn dialogue settings. We show that conversational systems that leverage DialoGPT generate more relevant, contentful and context-consistent responses than strong baseline systems. The pre-trained model and training pipeline are publicly released to facilitate research into neural response generation and the development of more intelligent open-domain dialogue systems.

Label Noise in Context [Website][PDF]

Michael Desmond, Catherine Finegan-Dollak, Jeff Boston, and Matt Arnold

Label noise—incorrectly or ambiguously labeled training examples—can negatively impact model performance. Although noise detection techniques have been around for decades, practitioners rarely apply them, as manual noise remediation is a tedious process. Examples incorrectly flagged as noise waste reviewers' time, and correcting label noise without guidance can be difficult. We propose LNIC, a noise-detection method that uses an example's neighborhood within the training set to (a) reduce false positives and (b) provide an explanation as to why the ex- ample was flagged as noise. We demonstrate on several short-text classification datasets that LNIC outperforms the state of the art on measures of precision and F0.5-score. We also show how LNIC's training set context helps a reviewer to understand and correct label noise in a dataset. The LNIC tool lowers the barriers to label noise remediation, increasing its utility for NLP practitioners.

Photon: A Robust Cross-Domain Text-to-SQL System [Website][PDF]

Jichuan Zeng, Xi Victoria Lin, Steven C.H. Hoi, Richard Socher, Caiming Xiong, Michael Lyu, and Irwin King

Natural language interfaces to databases(NLIDB) democratize end user access to relational data. Due to fundamental differences between natural language communication and programming, it is common for end users to issue questions that are ambiguous to the system or fall outside the semantic scope of its underlying query language. We present PHOTON, a robust, modular, cross-domain NLIDB that can flag natural language input to which a SQL mapping cannot be immediately determined. PHOTON consists of a strong neural semantic parser (63.2% structure accuracy on the Spider dev benchmark), a human-in-the-loop question corrector, a SQL executor and a response generator. The question corrector isa discriminative neural sequence editor which detects confusion span(s) in the input question and suggests rephrasing until a translatable input is given by the user or a maximum number of iterations are conducted. Experiments on simulated data show that the proposed method effectively improves the robustness of text-to-SQL system against untranslatable user input.The live demo of our system is available at http://www.naturalsql.com

Session 10A Overview – Tuesday, July 7, 2020 20:00–21:00

Track A *Computational Social Science and Social Media-8* Abstracts	Balancing Objectives in Counseling Conversations: Advancing Forwards or Looking Backwards *Zhang and Danescu-Niculescu-Mizil* [Website] [PDF]	Detecting Perceived Emotions in Hurricane Disasters *Desai, Caragea, and Li* [Website] [PDF]	Hierarchical Modeling for User Personality Prediction: The Role of Message-Level Attention *Lynn, Balasubramanian, and Schwartz* [Website] [PDF]	Masking Actor Information Leads to Fairer Political Claims Detection *Dayanik and Padó* [Website] [PDF]	Measuring Forecasting Skill from Text *Zong, Ritter, and Hovy* [Website] [PDF]
	Text and Causal Inference: A Review of Using Text to Remove Confounding from Causal Estimates *Keith, Jensen, and O'Connor* [Website] [PDF]	Text-Based Ideal Points *Vafa, Naidu, and Blei* [Website] [PDF]	Understanding the Language of Political Agreement and Disagreement in Legislative Texts *Davoodi, Waltenburg, and Goldwasser* [Website] [PDF]	When do Word Embeddings Accurately Reflect Surveys on our Beliefs About People? *Joseph and Morgan* [Website] [PDF]	"Who said it, and Why?" Provenance for Natural Language Claims *Zhang, Ives, and Roth* [Website] [PDF]
	Would you Rather? A New Benchmark for Learning Machine Alignment with Cultural Values and Social Preferences *Tay, Ong, Fu, Chan, Chen, Luu, and Pal* [Website] [PDF]				
Track B *Dialogue and Interactive Systems-12* Abstracts	CraftAssist Instruction Parsing: Semantic Parsing for a Voxel-World Assistant *Srinet, Jernite, Gray, and* [Website] [PDF]	Don't Say That! Making Inconsistent Dialogue Unlikely with Unlikelihood Training *Li, Roller, Kulikov, Welleck, Boureau, Cho, and Weston* [Website] [PDF]			
Track C *Interpretability and Analysis of Models for NLP-7* Abstracts	A Re-evaluation of Knowledge Graph Completion Methods *Sun, Vashishth, Sanyal, Talukdar, and Yang* [Website] [PDF]	Cross-Linguistic Syntactic Evaluation of Word Prediction Models *Mueller, Nicolai, Petrou-Zeniou, Talmina, and Linzen* [Website] [PDF]	ERASER: A Benchmark to Evaluate Rationalized NLP Models *DeYoung, Jain, Rajani, Lehman, Xiong, Socher, and Wallace* [Website] [PDF]	Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior? *Hase and Bansal* [Website] [PDF]	Finding Universal Grammatical Relations in Multilingual BERT *Chi, Hewitt, and Manning* [Website] [PDF]
	How does BERT's attention change when you fine-tune? An analysis methodology and a case study in negation scope *Zhao and Bethard* [Website] [PDF]	Influence Paths for Characterizing Subject-Verb Number Agreement in LSTM Language Models *Lu, Mardziel, Leino, Fredrikson, and Datta* [Website] [PDF]	Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings *Bommasani, Davis, and Cardie* [Website] [PDF]	Obtaining Faithful Interpretations from Compositional Neural Networks *Subramanian, Bogin, Gupta, Wolfson, Singh, Berant, and Gardner* [Website] [PDF]	Rationalizing Text Matching: Learning Sparse Alignments via Optimal Transport *Swanson, Yu, and Lei* [Website] [PDF]

	Spying on Your Neighbors: Fine-grained Probing of Contextual Embeddings for Information about Surrounding Words *Klafka and Ettinger* [Website][PDF]	[TACL] What BERT is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models *Ettinger* [Website][PDF]			
Track D *Question Answering-7* Abstracts	Benefits of Intermediate Annotations in Reading Comprehension *Dua, Singh, and Gardner* [Website][PDF]	Crossing Variational Autoencoders for Answer Retrieval *Yu, Wu, Zeng, Tao, Deng, and Jiang* [Website][PDF]	DeFormer: Decomposing Pre-trained Transformers for Faster Question Answering *Cao, Trivedi, Balasubramanian, and Balasubramanian* [Website][PDF]	[TACL] Investigating Prior Knowledge for Challenging Chinese Machine Reading Comprehension *Sun, Yu, Yu, and Cardie* [Website][PDF]	Logic-Guided Data Augmentation and Regularization for Consistent Question Answering *Asai and Hajishirzi* [Website][PDF]
	Probabilistic Assumptions Matter: Improved Models for Distantly-Supervised Document-Level Question Answering *Cheng, Chang, Lee, and Toutanova* [Website][PDF]	Selective Question Answering under Domain Shift *Kamath, Jia, and Liang* [Website][PDF]	[TACL] TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages *Clark, Palomaki, Nikolaev, Choi, Garrette, Collins, and Kwiatkowski* [Website][PDF]	Unsupervised Alignment-based Iterative Evidence Retrieval for Multi-hop Question Answering *Yadav, Bethard, and Surdeanu* [Website][PDF]	
Track E *Resources and Evaluation-11* Abstracts	Beyond Accuracy: Behavioral Testing of NLP Models with CheckList *Ribeiro, Wu, Guestrin, and Singh* [Website][PDF]	Code and Named Entity Recognition in StackOverflow *Tabassum, Maddela, Xu, and Ritter* [Website][PDF]	[CL] LINSPECTOR: Multilingual Probing Tasks for Word Representations *Şahin, Vania, Kuznetsov, and Gurevych* [Website][PDF]	[TACL] Paraphrase-Sense-Tagged Sentences *Cocos and Callison-Burch* [Website][PDF]	Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics *Mathur, Baldwin, and Cohn* [Website][PDF]
	Towards Holistic and Automatic Evaluation of Open-Domain Dialogue Generation *Pang, Nijkamp, Han, Zhou, Liu, and Tu* [Website][PDF]				
Track F *Sentiment Analysis, Stylistic Analysis, and Argument Mining-8* Abstracts	Agreement Prediction of Arguments in Cyber Argumentation for Detecting Stance Polarity and Intensity *Sirrianni, Liu, and Adams* [Website][PDF]	Cross-Lingual Unsupervised Sentiment Classification with Multi-View Transfer Learning *Fei and Li* [Website][PDF]	Efficient Pairwise Annotation of Argument Quality *Gienapp, Stein, Hagen, and Potthast* [Website][PDF]	Entity-Aware Dependency-Based Deep Graph Attention Network for Comparative Preference Classification *Ma, Mazumder, Wang, and Liu* [Website][PDF]	Modeling Label Semantics for Predicting Emotional Reactions *Gaonkar, Kwon, Bastan, Balasubramanian, and Chambers* [Website][PDF]

	OpinionDigest: A Simple Framework for Opinion Summarization *Suhara, Wang, Angelidis, and Tan* [Website][PDF]	A Comprehensive Analysis of Preprocessing for Word Representation Learning in Affective Tasks *Babanejad, Agrawal, An, and Papagelis* [Website][PDF]			
Track G *Theme-2* Abstracts	(Re)construing Meaning in NLP *Trott, Timponi Torrent, Chang, and Schneider* [Website][PDF]	Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data *Bender and Koller* [Website][PDF]	Examining Citations of Natural Language Processing Literature *Mohammad* [Website][PDF]	How Can We Accelerate Progress Towards Human-like Linguistic Generalization? *Linzen* [Website][PDF]	How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence *Zhong, Xiao, Tu, Zhang, Liu, and Sun* [Website][PDF]
	Intermediate-Task Transfer Learning with Pretrained Language Models: When and Why Does It Work? *Pruksachatkun, Phang, Liu, Htut, Zhang, Pang, Vania, Kann, and Bowman* [Website][PDF]	Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview *Shah, Schwartz, and Hovy* [Website][PDF]	What Does BERT with Vision Look At? *Li, Yatskar, Yin, Hsieh, and Chang* [Website][PDF]		

Session 10A Details

Session 10A: Computational Social Science and Social Media-8

Balancing Objectives in Counseling Conversations: Advancing Forwards or Looking Backwards [Website][PDF]
Justine Zhang and Cristian Danescu-Niculescu-Mizil 20:00–21:00

Throughout a conversation, participants make choices that can orient the flow of the interaction. Such choices are particularly salient in the consequential domain of crisis counseling, where a difficulty for counselors is balancing between two key objectives: advancing the conversation towards a resolution, and empathetically addressing the crisis situation. In this work, we develop an unsupervised methodology to quantify how counselors manage this balance. Our main intuition is that if an utterance can only receive a narrow range of appropriate replies, then its likely aim is to advance the conversation forwards, towards a target within that range. Likewise, an utterance that can only appropriately follow a narrow range of possible utterances is likely aimed backwards at addressing a specific situation within that range. By applying this intuition, we can map each utterance to a continuous orientation axis that captures the degree to which it is intended to direct the flow of the conversation forwards or backwards. This unsupervised method allows us to characterize counselor behaviors in a large dataset of crisis counseling conversations, where we show that known counseling strategies intuitively align with this axis. We also illustrate how our measure can be indicative of a conversation's progress, as well as its effectiveness.

Detecting Perceived Emotions in Hurricane Disasters [Website][PDF]
Shrey Desai, Cornelia Caragea, and Junyi Jessy Li 20:00–21:00

Natural disasters (e.g., hurricanes) affect millions of people each year, causing widespread destruction in their wake. People have recently taken to social media websites (e.g., Twitter) to share their sentiments and feelings with the larger community. Consequently, these platforms have become instrumental in understanding and perceiving emotions at scale. In this paper, we introduce HurricaneEmo, an emotion dataset of 15,000 English tweets spanning three hurricanes: Harvey, Irma, and Maria. We present a comprehensive study of fine-grained emotions and propose classification tasks to discriminate between coarse-grained emotion groups. Our best BERT model, even after task-guided pre-training which leverages unlabeled Twitter data, achieves only 68% accuracy (averaged across all groups). HurricaneEmo serves not only as a challenging benchmark for models but also as a valuable resource for analyzing emotions in disaster-centric domains.

Hierarchical Modeling for User Personality Prediction: The Role of Message-Level Attention [Website][PDF]
Veronica Lynn, Niranjan Balasubramanian, and H. Andrew Schwartz 20:00–21:00

Not all documents are equally important. Language processing is increasingly finding use as a supplement for questionnaires to assess psychological attributes of consenting individuals, but most approaches neglect to consider whether all documents of an individual are equally informative. In this paper, we present a novel model that uses message-level attention to learn the relative weight of users' social media posts for assessing their five factor personality traits. We demonstrate that models with message-level attention outperform those with word-level attention, and ultimately yield state-of-the-art accuracies for all five traits by using both word and message attention in combination with past approaches (an average increase in Pearson r of 2.5%). In addition, examination of the high-signal posts identified by our model provides insight into the relationship between language and personality, helping to inform future work.

Masking Actor Information Leads to Fairer Political Claims Detection [Website][PDF]
Erenay Dayanik and Sebastian Padó 20:00–21:00

A central concern in Computational Social Sciences (CSS) is fairness: where the role of NLP is to scale up text analysis to large corpora, the quality of automatic analyses should be as independent as possible of textual properties. We analyze the performance of a state-of-the-art neural model on the task of political claims detection (i.e., the identification of forward-looking statements made by political actors) and identify a strong frequency bias: claims made by frequent actors are recognized better. We propose two simple debiasing methods which mask proper names and pronouns during training of the model, thus removing personal information bias. We find that (a) these methods significantly decrease frequency bias while keeping the overall performance stable; and (b) the resulting models improve when evaluated in an out-of-domain setting.

Measuring Forecasting Skill from Text [Website][PDF]
Shi Zong, Alan Ritter, and Eduard Hovy 20:00–21:00

People vary in their ability to make accurate predictions about the future. Prior studies have shown that some individuals can predict the outcome of future events with consistently better accuracy. This leads to a natural question: what makes some forecasters better than others? In this paper we explore connections between the language people use to describe their predictions and their forecasting skill. Datasets from two different forecasting domains are explored: (1) geopolitical forecasts from Good Judgment Open, an online prediction forum and (2) a corpus of company earnings forecasts made by financial analysts. We present a number of linguistic metrics which are computed over text associated with people's predictions about the future including: uncertainty, readability, and emotion. By studying linguistic factors associated with predictions, we are able to shed some light on the approach taken by skilled forecasters. Furthermore, we demonstrate that it is possible to accurately predict forecasting skill using a model that

is based solely on language. This could potentially be useful for identifying accurate predictions or potentially skilled forecasters earlier.

Text and Causal Inference: A Review of Using Text to Remove Confounding from Causal Estimates [Website][PDF]
Katherine Keith, David Jensen, and Brendan O'Connor 20:00–21:00

Many applications of computational social science aim to infer causal conclusions from non-experimental data. Such observational data often contains confounders, variables that influence both potential causes and potential effects. Unmeasured or latent confounders can bias causal estimates, and this has motivated interest in measuring potential confounders from observed text. For example, an individual's entire history of social media posts or the content of a news article could provide a rich measurement of multiple confounders.Yet, methods and applications for this problem are scattered across different communities and evaluation practices are inconsistent.This review is the first to gather and categorize these examples and provide a guide to data-processing and evaluation decisions. Despite increased attention on adjusting for confounding using text, there are still many open problems, which we highlight in this paper.

Text-Based Ideal Points [Website][PDF]
Keyon Vafa, Suresh Naidu, and David Blei 20:00–21:00

Ideal point models analyze lawmakers' votes to quantify their political positions, or ideal points. But votes are not the only way to express a political position. Lawmakers also give speeches, release press statements, and post tweets. In this paper, we introduce the text-based ideal point model (TBIP), an unsupervised probabilistic topic model that analyzes texts to quantify the political positions of its authors. We demonstrate the TBIP with two types of politicized text data: U.S. Senate speeches and senator tweets. Though the model does not analyze their votes or political affiliations, the TBIP separates lawmakers by party, learns interpretable politicized topics, and infers ideal points close to the classical vote-based ideal points. One benefit of analyzing texts, as opposed to votes, is that the TBIP can estimate ideal points of anyone who authors political texts, including non-voting actors. To this end, we use it to study tweets from the 2020 Democratic presidential candidates. Using only the texts of their tweets, it identifies them along an interpretable progressive-to-moderate spectrum.

Understanding the Language of Political Agreement and Disagreement in Legislative Texts [Website][PDF]
Maryam Davoodi, Eric Waltenburg, and Dan Goldwasser 20:00–21:00

While national politics often receive the spotlight, the overwhelming majority of legislation proposed, discussed, and enacted is done at the state level. Despite this fact, there is little awareness of the dynamics that lead to adopting these policies. In this paper, we take the first step towards a better understanding of these processes and the underlying dynamics that shape them, using data-driven methods. We build a new large-scale dataset, from multiple data sources, connecting state bills and legislator information, geographical information about their districts, and donations and donors' information. We suggest a novel task, predicting the legislative body's vote breakdown for a given bill, according to different criteria of interest, such as gender, rural-urban and ideological splits. Finally, we suggest a shared relational embedding model, representing the interactions between the text of the bill and the legislative context in which it is presented. Our experiments show that providing this context helps improve the prediction over strong text-based models.

When do Word Embeddings Accurately Reflect Surveys on our Beliefs About People? [Website][PDF]
Kenneth Joseph and Jonathan Morgan 20:00–21:00

Social biases are encoded in word embeddings. This presents a unique opportunity to study society historically and at scale, and a unique danger when embeddings are used in downstream applications. Here, we investigate the extent to which publicly-available word embeddings accurately reflect beliefs about certain kinds of people as measured via traditional survey methods. We find that biases found in word embeddings do, on average, closely mirror survey data across seventeen dimensions of social meaning. However, we also find that biases in embeddings are much more reflective of survey data for some dimensions of meaning (e.g. gender) than others (e.g. race), and that we can be highly confident that embedding-based measures reflect survey data only for the most salient biases.

"Who said it, and Why?" Provenance for Natural Language Claims [Website][PDF]
Yi Zhang, Zachary Ives, and Dan Roth 20:00–21:00

In an era where generating content and publishing it is so easy, we are bombarded with information and are exposed to all kinds of claims, some of which do not always rank high on the truth scale. This paper suggests that the key to a longer-term, holistic, and systematic approach to navigating this information pollution is capturing the provenance of claims. To do that, we develop a formal definition of provenance graph for a given natural language claim, aiming to understand where the claim may come from and how it has evolved. To construct the graph, we model provenance inference, formulated mainly as an information extraction task and addressed via a textual entailment model. We evaluate our approach using two benchmark datasets, showing initial success in capturing the notion of provenance and its effectiveness on the application of claim verification.

Would you Rather? A New Benchmark for Learning Machine Alignment with Cultural Values and Social Preferences [Website][PDF]
Yi Tay, Donovan Ong, Jie Fu, Alvin Chan, Nancy Chen, Anh Tuan Luu, and Chris Pal 20:00–21:00

Understanding human preferences, along with cultural and social nuances, lives at the heart of natural language understanding. Concretely, we present a new task and corpus for learning alignments between machine and human preferences. Our newly introduced problem is concerned with predicting the preferable options from two sentences

describing scenarios that may involve social, cultural, ethical, or moral situations. Our problem is framed as a natural language inference task with crowd-sourced preference votes by human players, obtained from a gamified voting platform. Along with the release of a new dataset of 200K data points, we benchmark several state-of-the-art neural models, along with BERT and friends on this task. Our experimental results show that current state-of-the-art NLP models still leave much room for improvement.

Session 10A: Dialogue and Interactive Systems-12

CraftAssist Instruction Parsing: Semantic Parsing for a Voxel-World Assistant [Website][PDF]
Kavya Srinet, Yacine Jernite, Jonathan Gray, and arthur szlam arthur 20:00–21:00

We propose a semantic parsing dataset focused on instruction-driven communication with an agent in the game Minecraft. The dataset consists of 7K human utterances and their corresponding parses. Given proper world state, the parses can be interpreted and executed in game. We report the performance of baseline models, and analyze their successes and failures.

Don't Say That! Making Inconsistent Dialogue Unlikely with Unlikelihood Training [Website][PDF]
Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston 20:00–21:00

Generative dialogue models currently suffer from a number of problems which standard maximum likelihood training does not address. They tend to produce generations that (i) rely too much on copying from the context, (ii) contain repetitions within utterances, (iii) overuse frequent words, and (iv) at a deeper level, contain logical flaws.In this work we show how all of these problems can be addressed by extending the recently introduced unlikelihood loss (Welleck et al., 2019) to these cases. We show that appropriate loss functions which regularize generated outputs to match human distributions are effective for the first three issues. For the last important general issue, we show applying unlikelihood to collected data of what a model should not do is effective for improving logical consistency, potentially paving the way to generative models with greater reasoning ability. We demonstrate the efficacy of our approach across several dialogue tasks.

Session 10A: Interpretability and Analysis of Models for NLP-7

A Re-evaluation of Knowledge Graph Completion Methods [Website][PDF]
Zhiqing Sun, Shikhar Vashishth, Soumya Sanyal, Partha Talukdar, and Yiming Yang 20:00–21:00

Knowledge Graph Completion (KGC) aims at automatically predicting missing links for large-scale knowledge graphs. A vast number of state-of-the-art KGC techniques have got published at top conferences in several research fields, including data mining, machine learning, and natural language processing. However, we notice that several recent papers report very high performance, which largely outperforms previous state-of-the-art methods. In this paper, we find that this can be attributed to the inappropriate evaluation protocol used by them and propose a simple evaluation protocol to address this problem. The proposed protocol is robust to handle bias in the model, which can substantially affect the final results. We conduct extensive experiments and report performance of several existing methods using our protocol. The reproducible code has been made publicly available.

Cross-Linguistic Syntactic Evaluation of Word Prediction Models [Website][PDF]
Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen 20:00–21:00

A range of studies have concluded that neural word prediction models can distinguish grammatical from ungrammatical sentences with high accuracy. However, these studies are based primarily on monolingual evidence from English. To investigate how these models' ability to learn syntax varies by language, we introduce CLAMS (Cross-Linguistic Assessment of Models on Syntax), a syntactic evaluation suite for monolingual and multilingual models. CLAMS includes subject-verb agreement challenge sets for English, French, German, Hebrew and Russian, generated from grammars we develop. We use CLAMS to evaluate LSTM language models as well as monolingual and multilingual BERT. Across languages, monolingual LSTMs achieved high accuracy on dependencies without attractors, and generally poor accuracy on agreement across object relative clauses. On other constructions, agreement accuracy was generally higher in languages with richer morphology. Multilingual models generally underperformed monolingual models. Multilingual BERT showed high syntactic accuracy on English, but noticeable deficiencies in other languages.

ERASER: A Benchmark to Evaluate Rationalized NLP Models [Website][PDF]
Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace 20:00–21:00

State-of-the-art models in NLP are now predominantly based on deep neural networks that are opaque in terms of how they come to make predictions. This limitation has increased interest in designing more interpretable deep models for NLP that reveal the 'reasoning' behind model outputs. But work in this direction has been conducted on different datasets and tasks with correspondingly unique aims and metrics; this makes it difficult to track progress. We propose the **E**valuating **R**ationales **A**nd **S**imple **E**nglish **R**easoning (**ERASER** a benchmark to advance research on interpretable models in NLP. This benchmark comprises multiple datasets and tasks for which human annotations of "rationales" (supporting evidence) have been collected. We propose several metrics that aim to capture how well the rationales provided by models align with human rationales, and also how *faithful* these rationales are (i.e., the degree to which provided rationales influenced the corresponding predictions). Our hope is that releasing this benchmark facilitates progress on designing more interpretable NLP systems. The benchmark, code, and documentation are available at `https://www.eraserbenchmark.com/`

Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior? [Website][PDF]
Peter Hase and Mohit Bansal 20:00–21:00

Algorithmic approaches to interpreting machine learning models have proliferated in recent years. We carry out human subject tests that are the first of their kind to isolate the effect of algorithmic explanations on a key aspect of model interpretability, simulatability, while avoiding important confounding experimental factors. A model is simulatable when a person can predict its behavior on new inputs. Through two kinds of simulation tests involving text and tabular data, we evaluate five explanations methods: (1) LIME, (2) Anchor, (3) Decision Boundary, (4) a Prototype model, and (5) a Composite approach that combines explanations from each method. Clear evidence of method effectiveness is found in very few cases: LIME improves simulatability in tabular classification, and our Prototype method is effective in counterfactual simulation tests. We also collect subjective ratings of explanations, but we do not find that ratings are predictive of how helpful explanations are. Our results provide the first reliable and comprehensive estimates of how explanations influence simulatability across a variety of explanation methods and data domains. We show that (1) we need to be careful about the metrics we use to evaluate explanation methods, and (2) there is significant room for improvement in current methods.

Finding Universal Grammatical Relations in Multilingual BERT [Website][PDF]
Ethan A. Chi, John Hewitt, and Christopher D. Manning 20:00–21:00

Recent work has found evidence that Multilingual BERT (mBERT), a transformer-based multilingual masked language model, is capable of zero-shot cross-lingual transfer, suggesting that some aspects of its representations are shared cross-lingually. To better understand this overlap, we extend recent work on finding syntactic trees in neural networks' internal representations to the multilingual setting. We show that subspaces of mBERT representations recover syntactic tree distances in languages other than English, and that these subspaces are approximately shared across languages. Motivated by these results, we present an unsupervised analysis method that provides evidence mBERT learns representations of syntactic dependency labels, in the form of clusters which largely agree with the Universal Dependencies taxonomy. This evidence suggests that even without explicit supervision, multilingual masked language models learn certain linguistic universals.

How does BERT's attention change when you fine-tune? An analysis methodology and a case study in negation scope [Website][PDF]
Yiyun Zhao and Steven Bethard 20:00–21:00

Large pretrained language models like BERT, after fine-tuning to a downstream task, have achieved high performance on a variety of NLP problems. Yet explaining their decisions is difficult despite recent work probing their internal representations. We propose a procedure and analysis methods that take a hypothesis of how a transformer-based model might encode a linguistic phenomenon, and test the validity of that hypothesis based on a comparison between knowledge-related downstream tasks with downstream control tasks, and measurement of cross-dataset consistency. We apply this methodology to test BERT and RoBERTa on a hypothesis that some attention heads will consistently attend from a word in negation scope to the negation cue. We find that after fine-tuning BERT and RoBERTa on a negation scope task, the average attention head improves its sensitivity to negation and its attention consistency across negation datasets compared to the pre-trained models. However, only the base models (not the large models) improve compared to a control task, indicating there is evidence for a shallow encoding of negation only in the base models.

Influence Paths for Characterizing Subject-Verb Number Agreement in LSTM Language Models [Website][PDF]
Kaiji Lu, Piotr Mardziel, Klas Leino, Matt Fredrikson, and Anupam Datta 20:00–21:00

LSTM-based recurrent neural networks are the state-of-the-art for many natural language processing (NLP) tasks. Despite their performance, it is unclear whether, or how, LSTMs learn structural features of natural languages such as subject-verb number agreement in English. Lacking this understanding, the generality of LSTM performance on this task and their suitability for related tasks remains uncertain. Further, errors cannot be properly attributed to a lack of structural capability, training data omissions, or other exceptional faults. We introduce *influence paths*, a causal account of structural properties as carried by paths across gates and neurons of a recurrent neural network. The approach refines the notion of influence (the subject's grammatical number has influence on the grammatical number of the subsequent verb) into a set of gate or neuron-level paths. The set localizes and segments the concept (e.g., subject-verb agreement), its constituent elements (e.g., the subject), and related or interfering elements (e.g., attractors). We exemplify the methodology on a widely-studied multi-layer LSTM language model, demonstrating its accounting for subject-verb number agreement. The results offer both a finer and a more complete view of an LSTM's handling of this structural aspect of the English language than prior results based on diagnostic classifiers and ablation.

Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings [Website][PDF]
Rishi Bommasani, Kelly Davis, and Claire Cardie 20:00–21:00

Contextualized representations (e.g. ELMo, BERT) have become the default pretrained representations for downstream NLP applications. In some settings, this transition has rendered their static embedding predecessors (e.g. Word2Vec, GloVe) obsolete. As a side-effect, we observe that older interpretability methods for static embeddings — while more diverse and mature than those available for their dynamic counterparts — are underutilized in studying newer contextualized representations. Consequently, we introduce simple and fully general methods for converting from contextualized representations to static lookup-table embeddings which we apply to 5 popular pretrained models and 9 sets of pretrained weights. Our analysis of the resulting static embeddings notably reveals that pooling over many contexts significantly improves representational quality under intrinsic evaluation. Complementary to analyzing representational quality, we consider social biases encoded in pretrained representations with respect to gender, race/ethnicity, and religion and find that bias is encoded disparately across pretrained models and internal layers even for models with the same training data. Concerningly, we find dramatic inconsistencies between social bias estimators for word embeddings.

Obtaining Faithful Interpretations from Compositional Neural Networks [Website][PDF]
Sanjay Subramanian, Ben Bogin, Nitish Gupta, Tomer Wolfson, Sameer Singh, Jonathan Berant, and Matt Gardner 20:00–21:00

Neural module networks (NMNs) are a popular approach for modeling compositionality: they achieve high accuracy when applied to problems in language and vision, while reflecting the compositional structure of the problem in the network architecture. However, prior work implicitly assumed that the structure of the network modules, describing the abstract reasoning process, provides a faithful explanation of the model's reasoning; that is, that all modules perform their intended behaviour. In this work, we propose and conduct a systematic evaluation of the intermediate outputs of NMNs on NLVR2 and DROP, two datasets which require composing multiple reasoning steps. We find that the intermediate outputs differ from the expected output, illustrating that the network structure does not provide a faithful explanation of model behaviour. To remedy that, we train the model with auxiliary supervision and propose particular choices for module architecture that yield much better faithfulness, at a minimal cost to accuracy.

Rationalizing Text Matching: Learning Sparse Alignments via Optimal Transport [Website][PDF]
Kyle Swanson, Lili Yu, and Tao Lei 20:00–21:00

Selecting input features of top relevance has become a popular method for building self-explaining models. In this work, we extend this selective rationalization approach to text matching, where the goal is to jointly select and align text pieces, such as tokens or sentences, as a justification for the downstream prediction. Our approach employs optimal transport (OT) to find a minimal cost alignment between the inputs. However, directly applying OT often produces dense and therefore uninterpretable alignments. To overcome this limitation, we introduce novel constrained variants of the OT problem that result in highly sparse alignments with controllable sparsity. Our model is end-to-end differentiable using the Sinkhorn algorithm for OT and can be trained without any alignment annotations. We

evaluate our model on the StackExchange, MultiNews, e-SNLI, and MultiRC datasets. Our model achieves very sparse rationale selections with high fidelity while preserving prediction accuracy compared to strong attention baseline models.

Spying on Your Neighbors: Fine-grained Probing of Contextual Embeddings for Information about Surrounding Words [Website] [PDF]

Josef Klafka and Allyson Ettinger 20:00–21:00

Although models using contextual word embeddings have achieved state-of-the-art results on a host of NLP tasks, little is known about exactly what information these embeddings encode about the context words that they are understood to reflect. To address this question, we introduce a suite of probing tasks that enable fine-grained testing of contextual embeddings for encoding of information about surrounding words. We apply these tasks to examine the popular BERT, ELMo and GPT contextual encoders, and find that each of our tested information types is indeed encoded as contextual information across tokens, often with near-perfect recoverability—but the encoders vary in which features they distribute to which tokens, how nuanced their distributions are, and how robust the encoding of each feature is to distance. We discuss implications of these results for how different types of models break down and prioritize word-level context information when constructing token embeddings.

[TACL] What BERT is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models [Website] [PDF]

Allyson Ettinger 20:00–21:00

Pre-training by language modeling has become a popular and successful approach to NLP tasks, but we have yet to understand exactly what linguistic capacities these pre-training processes confer upon models. In this paper we introduce a suite of diagnostics drawn from human language experiments, which allow us to ask targeted questions about information used by language models for generating predictions in context. As a case study, we apply these diagnostics to the popular BERT model, finding that it can generally distinguish good from bad completions involving shared category or role reversal, albeit with less sensitivity than humans, and it robustly retrieves noun hypernyms, but it struggles with challenging inference and role-based event prediction – and in particular, it shows clear insensitivity to the contextual impacts of negation.

Session 10A: Question Answering-7

Benefits of Intermediate Annotations in Reading Comprehension [Website][PDF]
Dheeru Dua, Sameer Singh, and Matt Gardner 20:00–21:00

Complex compositional reading comprehension datasets require performing latent sequential decisions that are learned via supervision from the final answer. A large combinatorial space of possible decision paths that result in the same answer, compounded by the lack of intermediate supervision to help choose the right path, makes the learning particularly hard for this task. In this work, we study the benefits of collecting intermediate reasoning supervision along with the answer during data collection. We find that these intermediate annotations can provide two-fold benefits. First, we observe that for any collection budget, spending a fraction of it on intermediate annotations results in improved model performance, for two complex compositional datasets: DROP and Quoref. Second, these annotations encourage the model to learn the correct latent reasoning steps, helping combat some of the biases introduced during the data collection process.

Crossing Variational Autoencoders for Answer Retrieval [Website][PDF]
Wenhao Yu, Lingfei Wu, Qingkai Zeng, Shu Tao, Yu Deng, and Meng Jiang 20:00–21:00

Answer retrieval is to find the most aligned answer from a large set of candidates given a question. Learning vector representations of questions/answers is the key factor. Question-answer alignment and question/answer semantics are two important signals for learning the representations. Existing methods learned semantic representations with dual encoders or dual variational auto-encoders. The semantic information was learned from language models or question-to-question (answer-to-answer) generative processes. However, the alignment and semantics were too separate to capture the aligned semantics between question and answer. In this work, we propose to cross variational auto-encoders by generating questions with aligned answers and generating answers with aligned questions. Experiments show that our method outperforms the state-of-the-art answer retrieval method on SQuAD.

DeFormer: Decomposing Pre-trained Transformers for Faster Question Answering [Website][PDF]
Qingqing Cao, Harsh Trivedi, Aruna Balasubramanian, and Niranjan Balasubramanian 20:00–21:00

Transformer-based QA models use input-wide self-attention – i.e. across both the question and the input passage – at all layers, causing them to be slow and memory-intensive. It turns out that we can get by without input-wide self-attention at all layers, especially in the lower layers. We introduce DeFormer, a decomposed transformer, which substitutes the full self-attention with question-wide and passage-wide self-attentions in the lower layers. This allows for question-independent processing of the input text representations, which in turn enables pre-computing passage representations reducing runtime compute drastically. Furthermore, because DeFormer is largely similar to the original model, we can initialize DeFormer with the pre-training weights of a standard transformer, and directly fine-tune on the target QA dataset. We show DeFormer versions of BERT and XLNet can be used to speed up QA by over 4.3x and with simple distillation-based losses they incur only a 1% drop in accuracy. We open source the code at https://github.com/StonyBrookNLP/deformer.

[TACL] Investigating Prior Knowledge for Challenging Chinese Machine Reading Comprehension [Website][PDF]
Kai Sun, Dian Yu, Dong Yu, and Claire Cardie 20:00–21:00

Machine reading comprehension tasks require a machine reader to answer questions relevant to the given document. In this paper, we present the first free-form multiple-Choice Chinese machine reading Comprehension dataset (C^3), containing 13,369 documents (dialogues or more formally written mixed-genre texts) and their associated 19,577 multiple-choice free-form questions collected from Chinese-as-a-second-language examinations.

Logic-Guided Data Augmentation and Regularization for Consistent Question Answering [Website][PDF]
Akari Asai and Hannaneh Hajishirzi 20:00–21:00

Many natural language questions require qualitative, quantitative or logical comparisons between two entities or events. This paper addresses the problem of improving the accuracy and consistency of responses to comparison questions by integrating logic rules and neural models. Our method leverages logical and linguistic knowledge to augment labeled training data and then uses a consistency-based regularizer to train the model. Improving the global consistency of predictions, our approach achieves large improvements over previous methods in a variety of question answering (QA) tasks, including multiple-choice qualitative reasoning, cause-effect reasoning, and extractive machine reading comprehension. In particular, our method significantly improves the performance of RoBERTa-based models by 1-5% across datasets. We advance state of the art by around 5-8% on WIQA and QuaRel and reduce consistency violations by 58% on HotpotQA. We further demonstrate that our approach can learn effectively from limited data.

Probabilistic Assumptions Matter: Improved Models for Distantly-Supervised Document-Level Question Answering [Website][PDF]
Hao Cheng, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova 20:00–21:00

We address the problem of extractive question answering using document-level distant super-vision, pairing questions and relevant documents with answer strings. We compare previously used probability space and distant supervision assumptions (assumptions on the correspondence between the weak answer string labels and possible answer mention spans). We show that these assumptions interact, and that different configurations provide complementary benefits. We demonstrate that a multi-objective model can efficiently combine the advantages of multiple assump-

tions and outperform the best individual formulation. Our approach outperforms previous state-of-the-art models by 4.3 points in F1 on TriviaQA-Wiki and 1.7 points in Rouge-L on NarrativeQA summaries.

Selective Question Answering under Domain Shift [Website][PDF]
Amita Kamath, Robin Jia, and Percy Liang 20:00–21:00

To avoid giving wrong answers, question answering (QA) models need to know when to abstain from answering. Moreover, users often ask questions that diverge from the model's training data, making errors more likely and thus abstention more critical. In this work, we propose the setting of selective question answering under domain shift, in which a QA model is tested on a mixture of in-domain and out-of-domain data, and must answer (i.e., not abstain on) as many questions as possible while maintaining high accuracy. Abstention policies based solely on the model's softmax probabilities fare poorly, since models are overconfident on out-of-domain inputs. Instead, we train a calibrator to identify inputs on which the QA model errs, and abstain when it predicts an error is likely. Crucially, the calibrator benefits from observing the model's behavior on out-of-domain data, even if from a different domain than the test data. We combine this method with a SQuAD-trained QA model and evaluate on mixtures of SQuAD and five other QA datasets. Our method answers 56% of questions while maintaining 80% accuracy; in contrast, directly using the model's probabilities only answers 48% at 80% accuracy.

[TACL] TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages [Website][PDF]
Jonathan H Clark, Jennimaria Palomaki, Vitaly Nikolaev, Eunsol Choi, Dan Garrette, Michael Collins, and Tom Kwiatkowski 20:00–21:00

Confidently making progress on multilingual modeling requires challenging, trustworthy evaluations. We present TyDi QA, a question answering dataset covering 11 typologically diverse languages with 141K question-answer pairs. The languages of TyDi QA are diverse with regard to their typology — the set of linguistic features that each language expresses — such that we expect models performing well on this set to generalize across a large number of the languages in the world. We present a quantitative analysis of the data quality and example-level qualitative linguistic analyses of observed language phenomena that would not be found in English-only corpora. To provide a realistic information-seeking task and avoid priming effects, questions are written by people who want to know the answer, but don't know the answer yet, and the data is collected directly in each language without the use of translation. We provide initial quality measurements with a baseline model, suggesting a significant room for future work on this data.

Unsupervised Alignment-based Iterative Evidence Retrieval for Multi-hop Question Answering [Website][PDF]
Vikas Yadav, Steven Bethard, and Mihai Surdeanu 20:00–21:00

Evidence retrieval is a critical stage of question answering (QA), necessary not only to improve performance, but also to explain the decisions of the QA method. We introduce a simple, fast, and unsupervised iterative evidence retrieval method, which relies on three ideas: (a) an unsupervised alignment approach to soft-align questions and answers with justification sentences using only GloVe embeddings, (b) an iterative process that reformulates queries focusing on terms that are not covered by existing justifications, which (c) stops when the terms in the given question and candidate answers are covered by the retrieved justifications. Despite its simplicity, our approach outperforms all the previous methods (including supervised methods) on the evidence selection task on two datasets: MultiRC and QASC. When these evidence sentences are fed into a RoBERTa answer classification component, we achieve state-of-the-art QA performance on these two datasets.

Session 10A: Resources and Evaluation-11

Beyond Accuracy: Behavioral Testing of NLP Models with CheckList [Website][PDF]
Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh 20:00–21:00

Although measuring held-out accuracy has been the primary approach to evaluate generalization, it often overestimates the performance of NLP models, while alternative approaches for evaluating models either focus on individual tasks or on specific behaviors. Inspired by principles of behavioral testing in software engineering, we introduce CheckList, a task-agnostic methodology for testing NLP models. CheckList includes a matrix of general linguistic capabilities and test types that facilitate comprehensive test ideation, as well as a software tool to generate a large and diverse number of test cases quickly. We illustrate the utility of CheckList with tests for three tasks, identifying critical failures in both commercial and state-of-art models. In a user study, a team responsible for a commercial sentiment analysis model found new and actionable bugs in an extensively tested model. In another user study, NLP practitioners with CheckList created twice as many tests, and found almost three times as many bugs as users without it.

Code and Named Entity Recognition in StackOverflow [Website][PDF]
Jeniya Tabassum, Mounica Maddela, Wei Xu, and Alan Ritter 20:00–21:00

There is an increasing interest in studying natural language and computer code together, as large corpora of programming texts become readily available on the Internet. For example, StackOverflow currently has over 15 million programming related questions written by 8.5 million users. Meanwhile, there is still a lack of fundamental NLP techniques for identifying code tokens or software-related named entities that appear within natural language sentences. In this paper, we introduce a new named entity recognition (NER) corpus for the computer programming domain, consisting of 15,372 sentences annotated with 20 fine-grained entity types. We trained in-domain BERT representations (BERTOverflow) on 152 million sentences from StackOverflow, which lead to an absolute increase of +10 F_1 score over off-the-shelf BERT. We also present the SoftNER model which achieves an overall 79.10 F-1 score for code and named entity recognition on StackOverflow data. Our SoftNER model incorporates a context-independent code token classifier with corpus-level features to improve the BERT-based tagging model. Our code and data are available at: https://github.com/jeniyat/StackOverflowNER/

[CL] LINSPECTOR: Multilingual Probing Tasks for Word Representations [Website][PDF]
Gözde Gül Şahin, Clara Vania, Ilia Kuznetsov, and Iryna Gurevych 20:00–21:00

Despite an ever-growing number of word representation models introduced for a large number of languages, there is a lack of a standardized technique to provide insights into what is captured by these models. Such insights would help the community to get an estimate of the downstream task performance, as well as to design more informed neural architectures, while avoiding extensive experimentation that requires substantial computational resources not all researchers have access to. A recent development in NLP is to use simple classification tasks, also called probing tasks, that test for a single linguistic feature such as part-of-speech. Existing studies mostly focus on exploring the linguistic information encoded by the continuous representations of English text. However, from a typological perspective the morphologically poor English is rather an outlier: The information encoded by the word order and function words in English is often stored on a subword, morphological level in other languages. To address this, we introduce 15 type-level probing tasks such as case marking, possession, word length, morphological tag count, and pseudoword identification for 24 languages. We present a reusable methodology for creation and evaluation of such tests in a multilingual setting, which is challenging because of a lack of resources, lower quality of tools, and differences among languages. We then present experiments on several diverse multilingual word embedding models, in which we relate the probing task performance for a diverse set of languages to a range of five classic NLP tasks: POS-tagging, dependency parsing, semantic role labeling, named entity recognition, and natural language inference. We find that a number of probing tests have significantly high positive correlation to the downstream tasks, especially for morphologically rich languages. We show that our test scan be used to explore word embeddings or black-box neural models for linguistic cues in a multilingual setting. We release the probing data sets and the evaluation suite LINSPECTOR with https://github.com/UKPLab/linspector.

[TACL] Paraphrase-Sense-Tagged Sentences [Website][PDF]
Anne Cocos and Chris Callison-Burch 20:00–21:00

Many natural language processing tasks require discriminating the particular meaning of a word in context, but building corpora for developing sense-aware models can be a challenge. We present a large resource of example usages for words having a particular meaning, called Paraphrase-Sense-Tagged Sentences (PSTS). Built upon the premise that a word's paraphrases instantiate its fine-grained meanings – i.e. 'bug' has different meanings corresponding to its paraphrases 'fly' and 'microbe' – the resource contains up to 10,000 sentences for each of 3 million target-paraphrase pairs where the target word takes on the meaning of the paraphrase. We describe an automatic method based on bilingual pivoting used to enumerate sentences for PSTS, and present two models for ranking PSTS sentences based on their quality. Finally, we demonstrate the utility of PSTS by using it to build a dataset for the task of hypernym prediction in context. Training a model on this automatically-generated dataset produces accuracy that is competitive with a model trained on smaller datasets crafted with some manual effort.

Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics [Website][PDF]
Nitika Mathur, Timothy Baldwin, and Trevor Cohn 20:00–21:00

Automatic metrics are fundamental for the development and evaluation of machine translation systems. Judging whether, and to what extent, automatic metrics concur with the gold standard of human evaluation is not a straightforward problem. We show that current methods for judging metrics are highly sensitive to the translations used for

assessment, particularly the presence of outliers, which often leads to falsely confident conclusions about a metric's efficacy. Finally, we turn to pairwise system ranking, developing a method for thresholding performance improvement under an automatic metric against human judgements, which allows quantification of type I versus type II errors incurred, i.e., insignificant human differences in system quality that are accepted, and significant human differences that are rejected. Together, these findings suggest improvements to the protocols for metric evaluation and system performance evaluation in machine translation.

Towards Holistic and Automatic Evaluation of Open-Domain Dialogue Generation [Website] [PDF]
Bo Pang, Erik Nijkamp, Wenjuan Han, Linqi Zhou, Yixian Liu, and Kewei Tu 20:00–21:00

Open-domain dialogue generation has gained increasing attention in Natural Language Processing. Its evaluation requires a holistic means. Human ratings are deemed as the gold standard. As human evaluation is inefficient and costly, an automated substitute is highly desirable. In this paper, we propose holistic evaluation metrics that capture different aspects of open-domain dialogues. Our metrics consist of (1) GPT-2 based context coherence between sentences in a dialogue, (2) GPT-2 based fluency in phrasing, (3) n-gram based diversity in responses to augmented queries, and (4) textual-entailment-inference based logical self-consistency. The empirical validity of our metrics is demonstrated by strong correlations with human judgments. We open source the code and relevant materials.

Session 10A: Sentiment Analysis, Stylistic Analysis, and Argument Mining-8

Agreement Prediction of Arguments in Cyber Argumentation for Detecting Stance Polarity and Intensity [Website][PDF]
Joseph Sirrianni, Xiaoqing Liu, and Douglas Adams 20:00–21:00

In online debates, users express different levels of agreement/disagreement with one another's arguments and ideas. Often levels of agreement/disagreement are implicit in the text, and must be predicted to analyze collective opinions. Existing stance detection methods predict the polarity of a post's stance toward a topic or post, but don't consider the stance's degree of intensity. We introduce a new research problem, stance polarity and intensity prediction in response relationships between posts. This problem is challenging because differences in stance intensity are often subtle and require nuanced language understanding. Cyber argumentation research has shown that incorporating both stance polarity and intensity data in online debates leads to better discussion analysis. We explore five different learning models: Ridge-M regression, Ridge-S regression, SVR-RF-R, pkudblab-PIP, and T-PAN-PIP for predicting stance polarity and intensity in argumentation. These models are evaluated using a new dataset for stance polarity and intensity prediction collected using a cyber argumentation platform. The SVR-RF-R model performs best for prediction of stance polarity with an accuracy of 70.43% and intensity with RMSE of 0.596. This work is the first to train models for predicting a post's stance polarity and intensity in one combined value in cyber argumentation with reasonably good accuracy.

Cross-Lingual Unsupervised Sentiment Classification with Multi-View Transfer Learning [Website][PDF]
Hongliang Fei and Ping Li 20:00–21:00

Recent neural network models have achieved impressive performance on sentiment classification in English as well as other languages. Their success heavily depends on the availability of a large amount of labeled data or parallel corpus. In this paper, we investigate an extreme scenario of cross-lingual sentiment classification, in which the low-resource language does not have any labels or parallel corpus. We propose an unsupervised cross-lingual sentiment classification model named multi-view encoder-classifier (MVEC) that leverages an unsupervised machine translation (UMT) system and a language discriminator. Unlike previous language model (LM) based fine-tuning approaches that adjust parameters solely based on the classification error on training data, we employ the encoder-decoder framework of a UMT as a regularization component on the shared network parameters. In particular, the cross-lingual encoder of our model learns a shared representation, which is effective for both reconstructing input sentences of two languages and generating more representative views from the input for classification. Extensive experiments on five language pairs verify that our model significantly outperforms other models for 8/11 sentiment classification tasks.

Efficient Pairwise Annotation of Argument Quality [Website][PDF]
Lukas Gienapp, Benno Stein, Matthias Hagen, and Martin Potthast 20:00–21:00

We present an efficient annotation framework for argument quality, a feature difficult to be measured reliably as per previous work. A stochastic transitivity model is combined with an effective sampling strategy to infer high-quality labels with low effort from crowdsourced pairwise judgments. The model's capabilities are showcased by compiling Webis-ArgQuality-20, an argument quality corpus that comprises scores for rhetorical, logical, dialectical, and overall quality inferred from a total of 41,859 pairwise judgments among 1,271 arguments. With up to 93% cost savings, our approach significantly outperforms existing annotation procedures. Furthermore, novel insight into argument quality is provided through statistical analysis, and a new aggregation method to infer overall quality from individual quality dimensions is proposed.

Entity-Aware Dependency-Based Deep Graph Attention Network for Comparative Preference Classification [Website][PDF]
Nianzu Ma, Sahisnu Mazumder, Hao Wang, and Bing Liu 20:00–21:00

This paper studies the task of comparative preference classification (CPC). Given two entities in a sentence, our goal is to classify whether the first (or the second) entity is preferred over the other or no comparison is expressed at all between the two entities. Existing works either do not learn entity-aware representations well and fail to deal with sentences involving multiple entity pairs or use sequential modeling approaches that are unable to capture long-range dependencies between the entities. Some also use traditional machine learning approaches that do not generalize well. This paper proposes a novel Entity-aware Dependency-based Deep Graph Attention Network (ED-GAT) that employs a multi-hop graph attention over a dependency graph sentence representation to leverage both the semantic information from word embeddings and the syntactic information from the dependency graph to solve the problem. Empirical evaluation shows that the proposed model achieves the state-of-the-art performance in comparative preference classification.

Modeling Label Semantics for Predicting Emotional Reactions [Website][PDF]
Radhika Gaonkar, Heeyoung Kwon, Mohaddeseh Bastan, Niranjan Balasubramanian, and Nathanael Chambers 20:00–21:00

Predicting how events induce emotions in the characters of a story is typically seen as a standard multi-label classification task, which usually treats labels as anonymous classes to predict. They ignore information that may be conveyed by the emotion labels themselves. We propose that the semantics of emotion labels can guide a model's attention when representing the input story. Further, we observe that the emotions evoked by an event are often related: an event that evokes joy is unlikely to also evoke sadness. In this work, we explicitly model label classes via label embeddings, and add mechanisms that track label-label correlations both during training and inference. We

also introduce a new semi-supervision strategy that regularizes for the correlations on unlabeled data. Our empirical evaluations show that modeling label semantics yields consistent benefits, and we advance the state-of-the-art on an emotion inference task.

OpinionDigest: A Simple Framework for Opinion Summarization [Website][PDF]
Yoshihiko Suhara, Xiaolan Wang, Stefanos Angelidis, and Wang-Chiew Tan 20:00–21:00

We present OpinionDigest, an abstractive opinion summarization framework, which does not rely on gold-standard summaries for training. The framework uses an Aspect-based Sentiment Analysis model to extract opinion phrases from reviews, and trains a Transformer model to reconstruct the original reviews from these extractions. At summarization time, we merge extractions from multiple reviews and select the most popular ones. The selected opinions are used as input to the trained Transformer model, which verbalizes them into an opinion summary. OpinionDigest can also generate customized summaries, tailored to specific user needs, by filtering the selected opinions according to their aspect and/or sentiment. Automatic evaluation on Yelp data shows that our framework outperforms competitive baselines. Human studies on two corpora verify that OpinionDigest produces informative summaries and shows promising customization capabilities.

A Comprehensive Analysis of Preprocessing for Word Representation Learning in Affective Tasks [Website][PDF]
Nastaran Babanejad, Ameeta Agrawal, Aijun An, and Manos Papagelis 20:00–21:00

Affective tasks such as sentiment analysis, emotion classification, and sarcasm detection have been popular in recent years due to an abundance of user-generated data, accurate computational linguistic models, and a broad range of relevant applications in various domains. At the same time, many studies have highlighted the importance of text preprocessing, as an integral step to any natural language processing prediction model and downstream task. While preprocessing in affective systems is well-studied, preprocessing in word vector-based models applied to affective systems, is not. To address this limitation, we conduct a comprehensive analysis of the role of preprocessing techniques in affective analysis based on word vector models. Our analysis is the first of its kind and provides useful insights of the importance of each preprocessing technique when applied at the training phase, commonly ignored in pretrained word vector models, and/or at the downstream task phase.

Session 10A: Theme-2

(Re)construing Meaning in NLP [Website][PDF]
Sean Trott, Tiago Timponi Torrent, Nancy Chang, and Nathan Schneider 20:00–21:00

Human speakers have an extensive toolkit of ways to express themselves. In this paper, we engage with an idea largely absent from discussions of meaning in natural language understanding—namely, that the way something is expressed reflects different ways of conceptualizing or construing the information being conveyed. We first define this phenomenon more precisely, drawing on considerable prior work in theoretical cognitive semantics and psycholinguistics. We then survey some dimensions of construed meaning and show how insights from construal could inform theoretical and practical work in NLP.

Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data [Website][PDF]
Emily M. Bender and Alexander Koller 20:00–21:00

The success of the large neural language models on many NLP tasks is exciting. However, we find that these successes sometimes lead to hype in which these models are being described as "understanding" language or capturing "meaning". In this position paper, we argue that a system trained only on form has a priori no way to learn meaning. In keeping with the ACL 2020 theme of "Taking Stock of Where We've Been and Where We're Going", we argue that a clear understanding of the distinction between form and meaning will help guide the field towards better science around natural language understanding.

Examining Citations of Natural Language Processing Literature [Website][PDF]
Saif M. Mohammad 20:00–21:00

We extracted information from the ACL Anthology (AA) and Google Scholar (GS) to examine trends in citations of NLP papers. We explore questions such as: how well cited are papers of different types (journal articles, conference papers, demo papers, etc.)? how well cited are papers from different areas of within NLP? etc. Notably, we show that only about 56% of the papers in AA are cited ten or more times. CL Journal has the most cited papers, but its citation dominance has lessened in recent years. On average, long papers get almost three times as many citations as short papers; and papers on sentiment classification, anaphora resolution, and entity recognition have the highest median citations. The analyses presented here, and the associated dataset of NLP papers mapped to citations, have a number of uses including: understanding how the field is growing and quantifying the impact of different types of papers.

How Can We Accelerate Progress Towards Human-like Linguistic Generalization? [Website][PDF]
Tal Linzen 20:00–21:00

This position paper describes and critiques the Pretraining-Agnostic Identically Distributed (PAID) evaluation paradigm, which has become a central tool for measuring progress in natural language understanding. This paradigm consists of three stages: (1) pre-training of a word prediction model on a corpus of arbitrary size; (2) fine-tuning (transfer learning) on a training set representing a classification task; (3) evaluation on a test set drawn from the same distribution as that training set. This paradigm favors simple, low-bias architectures, which, first, can be scaled to process vast amounts of data, and second, can capture the fine-grained statistical properties of a particular data set, regardless of whether those properties are likely to generalize to examples of the task outside the data set. This contrasts with humans, who learn language from several orders of magnitude less data than the systems favored by this evaluation paradigm, and generalize to new tasks in a consistent way. We advocate for supplementing or replacing PAID with paradigms that reward architectures that generalize as quickly and robustly as humans.

How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence [Website][PDF]
Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun 20:00–21:00

Legal Artificial Intelligence (LegalAI) focuses on applying the technology of artificial intelligence, especially natural language processing, to benefit tasks in the legal domain. In recent years, LegalAI has drawn increasing attention rapidly from both AI researchers and legal professionals, as LegalAI is beneficial to the legal system for liberating legal professionals from a maze of paperwork. Legal professionals often think about how to solve tasks from rule-based and symbol-based methods, while NLP researchers concentrate more on data-driven and embedding methods. In this paper, we introduce the history, the current state, and the future directions of research in LegalAI. We illustrate the tasks from the perspectives of legal professionals and NLP researchers and show several representative applications in LegalAI. We conduct experiments and provide an in-depth analysis of the advantages and disadvantages of existing works to explore possible future directions. You can find the implementation of our work from https://github.com/thunlp/CLAIM.

Intermediate-Task Transfer Learning with Pretrained Language Models: When and Why Does It Work? [Website][PDF]
Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman 20:00–21:00

While pretrained models such as BERT have shown large gains across natural language understanding tasks, their performance can be improved by further training the model on a data-rich intermediate task, before fine-tuning it on a target task. However, it is still poorly understood when and why intermediate-task training is beneficial for a given target task. To investigate this, we perform a large-scale study on the pretrained RoBERTa model with 110 intermediate-target task combinations. We further evaluate all trained models with 25 probing tasks meant to reveal the specific skills that drive transfer. We observe that intermediate tasks requiring high-level inference and reasoning abilities tend to work best. We also observe that target task performance is strongly correlated with higher-level

abilities such as coreference resolution. However, we fail to observe more granular correlations between probing and target task performance, highlighting the need for further work on broad-coverage probing benchmarks. We also observe evidence that the forgetting of knowledge learned during pretraining may limit our analysis, highlighting the need for further work on transfer learning methods in these settings.

Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview [Website][PDF]
Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy 20:00–21:00

An increasing number of natural language processing papers address the effect of bias on predictions, introducing mitigation techniques at different parts of the standard NLP pipeline (data and models). However, these works have been conducted individually, without a unifying framework to organize efforts within the field. This situation leads to repetitive approaches, and focuses overly on bias symptoms/effects, rather than on their origins, which could limit the development of effective countermeasures. In this paper, we propose a unifying predictive bias framework for NLP. We summarize the NLP literature and suggest general mathematical definitions of predictive bias. We differentiate two consequences of bias: outcome disparities and error disparities, as well as four potential origins of biases: label bias, selection bias, model overamplification, and semantic bias. Our framework serves as an overview of predictive bias in NLP, integrating existing work into a single structure, and providing a conceptual baseline for improved frameworks.

What Does BERT with Vision Look At? [Website][PDF]
Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang 20:00–21:00

Pre-trained visually grounded language models such as ViLBERT, LXMERT, and UNITER have achieved significant performance improvement on vision-and-language tasks but what they learn during pre-training remains unclear. In this work, we demonstrate that certain attention heads of a visually grounded language model actively ground elements of language to image regions. Specifically, some heads can map entities to image regions, performing the task known as entity grounding. Some heads can even detect the syntactic relations between non-entity words and image regions, tracking, for example, associations between verbs and regions corresponding to their arguments. We denote this ability as *syntactic grounding*. We verify grounding both quantitatively and qualitatively, using Flickr30K Entities as a testbed.

Demo Session 5B

Time: *20:45–21:30*

Interactive Task Learning from GUI-Grounded Natural Language Instructions and Demonstrations [Website][PDF]
Toby Jia-Jun Li, Tom Mitchell, and Brad Myers

We show SUGILITE, an intelligent task automation agent that can learn new tasks and relevant associated concepts interactively from the user's natural language instructions and demonstrations, using the graphical user interfaces (GUIs) of third-party mobile apps. This system provides several interesting features: (1) it allows users to teach new task procedures and concepts through verbal instructions together with demonstration of the steps of a script using GUIs; (2) it supports users in clarifying their intents for demonstrated actions using GUI-grounded verbal instructions; (3) it infers parameters of tasks and their possible values in utterances using the hierarchical structures of the underlying app GUIs; and (4) it generalizes taught concepts to different contexts and task domains. We describe the architecture of the SUGILITE system, explain the design and implementation of its key features, and show a prototype in the form of a conversational assistant on Android.

exBERT: A Visual Analysis Tool to Explore Learned Representations in Transformer Models [Website][PDF]
Benjamin Hoover, Hendrik Strobelt, and Sebastian Gehrmann

Large Transformer-based language models can route and reshape complex information via their multi-headed attention mechanism. Although the attention never receives explicit supervision, it can exhibit recognizable patterns following linguistic or positional information. Analyzing the learned representations and attentions is paramount to furthering our understanding of the inner workings of these models. However, analyses have to catch up with the rapid release of new models and the growing diversity of investigation techniques. To support analysis for a wide variety of models, we introduce exBERT, a tool to help humans conduct flexible, interactive investigations and formulate hypotheses for the model-internal reasoning process. exBERT provides insights into the meaning of the contextual representations and attention by matching a human-specified input to similar contexts in large annotated datasets. By aggregating the annotations of the matched contexts, exBERT can quickly replicate findings from literature and extend them to previously not analyzed models.

Session 10B Overview – Tuesday, July 7, 2020 21:00–22:00

Track A *Discourse and Pragmatics-5* Abstracts	Discourse as a Function of Event: Profiling Discourse Structure in News Articles around the Main Event *Choubey, Lee, Huang, and Wang* [Website][PDF]	Harnessing the linguistic signal to predict scalar inferences *Schuster, Chen, and Degen* [Website][PDF]	Implicit Discourse Relation Classification: We Need to Talk about Evaluation *Kim, Feng, Gunasekara, and Lastras* [Website][PDF]	PeTra: A Sparsely Supervised Memory Model for People Tracking *Toshniwal, Ettinger, Gimpel, and Livescu* [Website][PDF]	ZPR2: Joint Zero Pronoun Recovery and Resolution using Multi-Task Learning and BERT *Song, Xu, Zhang, Chen, and Yu* [Website][PDF]
Track B *Ethics and NLP-5* Abstracts	Contextualizing Hate Speech Classifiers with Post-hoc Explanation *Kennedy, Jin, Mostafazadeh Davani, Dehghani, and Ren* [Website][PDF]	Double-Hard Debias: Tailoring Word Embeddings for Gender Bias Mitigation *Wang, Lin, Rajani, McCann, Ordonez, and Xiong* [Website][PDF]	Gender Bias in Multilingual Embeddings and Cross-Lingual Transfer *Zhao, Mukherjee, Hosseini, Chang, and Hassan Awadallah* [Website][PDF]	Language (Technology) is Power: A Critical Survey of "Bias" in NLP *Blodgett, Barocas, Daumé III, and Wallach* [Website][PDF]	Mitigating Gender Bias Amplification in Distribution by Posterior Regularization *Jia, Meng, Zhao, and Chang* [Website][PDF]
	Social Bias Frames: Reasoning about Social and Power Implications of Language *Sap, Gabriel, Qin, Jurafsky, Smith, and Choi* [Website][PDF]	Social Biases in NLP Models as Barriers for Persons with Disabilities *Hutchinson, Prabhakaran, Denton, Webster, Zhong, and Denuyl* [Website][PDF]	Toward Gender-Inclusive Coreference Resolution *Cao and Daumé III* [Website][PDF]	Towards Debiasing Sentence Representations *Liang, Li, Zheng, Lim, Salakhutdinov, and Morency* [Website][PDF]	
Track C *Interpretability and Analysis of Models for NLP-8* Abstracts	Explaining Black Box Predictions and Unveiling Data Artifacts through Influence Functions *Han, Wallace, and Tsvetkov* [Website][PDF]	Generating Hierarchical Explanations on Text Classification via Feature Interaction Detection *Chen, Zheng, and Ji* [Website][PDF]	Learning to Deceive with Attention-Based Explanations *Pruthi, Gupta, Dhingra, Neubig, and Lipton* [Website][PDF]	Learning to Faithfully Rationalize by Construction *Jain, Wiegreffe, Pinter, and Wallace* [Website][PDF]	On the Spontaneous Emergence of Discrete and Compositional Signals *Geffen Lan, Chemla, and Steinert-Threlkeld* [Website][PDF]
Track D *Language Grounding to Vision, Robotics and Beyond-4* Abstracts	Dense-Caption Matching and Frame-Selection Gating for Temporal Localization in VideoQA *Kim, Tang, and Bansal* [Website][PDF]	Shaping Visual Representations with Language for Few-Shot Classification *Mu, Liang, and Goodman* [Website][PDF]			
Track E *Machine Learning for NLP-11* Abstracts	Discrete Latent Variable Representations for Low-Resource Text Classification *Jin, Wiseman, Stratos, and Livescu* [Website][PDF]	Improving Transformer Models by Reordering their Sublayers *Press, Smith, and Levy* [Website][PDF]	Learning Constraints for Structured Prediction Using Rectifier Networks *Pan, Mehta, and Srikumar* [Website][PDF]	Pretraining with Contrastive Sentence Objectives Improves Discourse Performance of Language Models *Iter, Guu, Lansing, and Jurafsky* [Website][PDF]	SAFER: A Structure-free Approach for Certified Robustness to Adversarial Word Substitutions *Ye, Gong, and Liu* [Website][PDF]

	[TACL] SpanBERT: Improving Pre-training by Representing and Predicting Spans *Joshi, Chen, Liu, Weld, Zettlemoyer, and Levy* [Website][PDF]				
Track F *Question Answering-8* Abstracts	Clinical Reading Comprehension: A Thorough Analysis of the emrQA Dataset *Yue, Jimenez Gutierrez, and Sun* [Website][PDF]	On the Importance of Diversity in Question Generation for QA *Sultan, Chandel, Fernandez Astudillo, and Castelli* [Website][PDF]	SCDE: Sentence Cloze Dataset with High Quality Distractors From Examinations *Kong, Gangal, and Hovy* [Website][PDF]	The Cascade Transformer: an Application for Efficient Answer Sentence Selection *Soldaini and Moschitti* [Website][PDF]	Transformers to Learn Hierarchical Contexts in Multiparty Dialogue for Span-based Question Answering *Li and Choi* [Website][PDF]
Track G *Resources and Evaluation-12* Abstracts	A Recipe for Creating Multimodal Aligned Datasets for Sequential Tasks *Lin, Rao, Celikyilmaz, Nouri, Brockett, Dey, and Dolan* [Website][PDF]	Adversarial NLI: A New Benchmark for Natural Language Understanding *Nie, Williams, Dinan, Bansal, Weston, and Kiela* [Website][PDF]	Dialogue-Based Relation Extraction *Yu, Sun, Cardie, and Yu* [Website][PDF]	Dscorer: A Fast Evaluation Metric for Discourse Representation Structure Parsing *Liu, Cohen, and Lapata* [Website][PDF]	Facet-Aware Evaluation for Extractive Summarization *Mao, Liu, Zhu, Ren, and Han* [Website][PDF]
	More Diverse Dialogue Datasets via Diversity-Informed Data Collection *Stasaski, Yang, and Hearst* [Website][PDF]	Not All Claims are Created Equal: Choosing the Right Statistical Approach to Assess Hypotheses *Sadeqi Azer, Khashabi, Sabharwal, and Roth* [Website][PDF]	S2ORC: The Semantic Scholar Open Research Corpus *Lo, Wang, Neumann, Kinney, and Weld* [Website][PDF]	STARC: Structured Annotations for Reading Comprehension *Berzak, Malmaud, and Levy* [Website][PDF]	WinoWhy: A Deep Diagnosis of Essential Commonsense Knowledge for Answering Winograd Schema Challenge *Zhang, Zhao, and Song* [Website][PDF]
Track H *Speech and Multimodality-6* Abstracts	Towards end-2-end learning for predicting behavior codes from spoken utterances in psychotherapy conversations *Singla, Chen, Atkins, and Narayanan* [Website][PDF]				
Track I *Summarization-5* Abstracts	A Transformer-based Approach for Source Code Summarization *Ahmad, Chakraborty, Ray, and Chang* [Website][PDF]	Asking and Answering Questions to Evaluate the Factual Consistency of Summaries *Wang, Cho, and Lewis* [Website][PDF]	Discourse-Aware Neural Extractive Text Summarization *Xu, Gan, Cheng, and Liu* [Website][PDF]	Discrete Optimization for Unsupervised Sentence Summarization with Word-Level Extraction *Schumann, Mou, Lu, Vechtomova, and Markert* [Website][PDF]	Exploring Content Selection in Summarization of Novel Chapters *Ladhak, Li, Al-Onaizan, and McKeown* [Website][PDF]

FEQA: A Question Answering Evaluation Framework for Faithfulness Assessment in Abstractive Summarization *Durmus, He, and Diab* [Website][PDF]	Fact-based Content Weighting for Evaluating Abstractive Summarisation *Xu, Dušek, Li, Rieser, and Konstas* [Website][PDF]	Hooks in the Headline: Learning to Generate Headlines with Controlled Styles *Jin, Jin, Zhou, Orii, and Szolovits* [Website][PDF]	Knowledge Graph-Augmented Abstractive Summarization with Semantic-Driven Cloze Reward *Huang, Wu, and Wang* [Website][PDF]	Optimizing the Factual Correctness of a Summary: A Study of Summarizing Radiology Reports *Zhang, Merck, Tsai, Manning, and Langlotz* [Website][PDF]
Storytelling with Dialogue: A Critical Role Dungeons and Dragons Dataset *Rameshkumar and Bailey* [Website][PDF]	The Summary Loop: Learning to Write Abstractive Summaries Without Examples *Laban, Hsi, Canny, and Hearst* [Website][PDF]	Unsupervised Opinion Summarization as Copycat-Review Generation *Bražinskas, Lapata, and Titov* [Website][PDF]		

Session 10B Details

Session 10B: Discourse and Pragmatics-5

Discourse as a Function of Event: Profiling Discourse Structure in News Articles around the Main Event [Website][PDF]
Prafulla Kumar Choubey, Aaron Lee, Ruihong Huang, and Lu Wang 21:00–22:00

Understanding discourse structures of news articles is vital to effectively contextualize the occurrence of a news event. To enable computational modeling of news structures, we apply an existing theory of functional discourse structure for news articles that revolves around the main event and create a human-annotated corpus of 802 documents spanning over four domains and three media sources. Next, we propose several document-level neural-network models to automatically construct news content structures. Finally, we demonstrate that incorporating system predicted news structures yields new state-of-the-art performance for event coreference resolution. The news documents we annotated are openly available and the annotations are publicly released for future research.

Harnessing the linguistic signal to predict scalar inferences [Website][PDF]
Sebastian Schuster, Yuxing Chen, and Judith Degen 21:00–22:00

Pragmatic inferences often subtly depend on the presence or absence of linguistic features. For example, the presence of a partitive construction (of the) increases the strength of a so-called scalar inference: listeners perceive the inference that Chris did not eat all of the cookies to be stronger after hearing "Chris ate some of the cookies" than after hearing the same utterance without a partitive, "Chris ate some cookies". In this work, we explore to what extent neural network sentence encoders can learn to predict the strength of scalar inferences. We first show that an LSTM-based sentence encoder trained on an English dataset of human inference strength ratings is able to predict ratings with high accuracy ($r = 0.78$). We then probe the model's behavior using manually constructed minimal sentence pairs and corpus data. We first that the model inferred previously established associations between linguistic features and inference strength, suggesting that the model learns to use linguistic features to predict pragmatic inferences.

Implicit Discourse Relation Classification: We Need to Talk about Evaluation [Website][PDF]
Najoung Kim, Song Feng, Chulaka Gunasekara, and Luis Lastras 21:00–22:00

Implicit relation classification on Penn Discourse TreeBank (PDTB) 2.0 is a common benchmark task for evaluating the understanding of discourse relations. However, the lack of consistency in preprocessing and evaluation poses challenges to fair comparison of results in the literature. In this work, we highlight these inconsistencies and propose an improved evaluation protocol. Paired with this protocol, we report strong baseline results from pretrained sentence encoders, which set the new state-of-the-art for PDTB 2.0. Furthermore, this work is the first to explore fine-grained relation classification on PDTB 3.0. We expect our work to serve as a point of comparison for future work, and also as an initiative to discuss models of larger context and possible data augmentations for downstream transferability.

PeTra: A Sparsely Supervised Memory Model for People Tracking [Website][PDF]
Shubham Toshniwal, Allyson Ettinger, Kevin Gimpel, and Karen Livescu 21:00–22:00

We propose PeTra, a memory-augmented neural network designed to track entities in its memory slots. PeTra is trained using sparse annotation from the GAP pronoun resolution dataset and outperforms a prior memory model on the task while using a simpler architecture. We empirically compare key modeling choices, finding that we can simplify several aspects of the design of the memory module while retaining strong performance. To measure the people tracking capability of memory models, we (a) propose a new diagnostic evaluation based on counting the number of unique entities in text, and (b) conduct a small scale human evaluation to compare evidence of people tracking in the memory logs of PeTra relative to a previous approach. PeTra is highly effective in both evaluations, demonstrating its ability to track people in its memory despite being trained with limited annotation.

ZPR2: Joint Zero Pronoun Recovery and Resolution using Multi-Task Learning and BERT [Website][PDF]
Linfeng Song, Kun Xu, Yue Zhang, Jianshu Chen, and Dong Yu 21:00–22:00

Zero pronoun recovery and resolution aim at recovering the dropped pronoun and pointing out its anaphoric mentions, respectively. We propose to better explore their interaction by solving both tasks together, while the previous work treats them separately. For zero pronoun resolution, we study this task in a more realistic setting, where no parsing trees or only automatic trees are available, while most previous work assumes gold trees. Experiments on two benchmarks show that joint modeling significantly outperforms our baseline that already beats the previous state of the arts.

Session 10B: Ethics and NLP-5

Contextualizing Hate Speech Classifiers with Post-hoc Explanation [Website][PDF]
Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren 21:00–22:00

Hate speech classifiers trained on imbalanced datasets struggle to determine if group identifiers like "gay" or "black" are used in offensive or prejudiced ways. Such biases manifest in false positives when these identifiers are present, due to models' inability to learn the contexts which constitute a hateful usage of identifiers. We extract post-hoc explanations from fine-tuned BERT classifiers to detect bias towards identity terms. Then, we propose a novel regularization technique based on these explanations that encourages models to learn from the context of group identifiers in addition to the identifiers themselves. Our approach improved over baselines in limiting false positives on out-of-domain data while maintaining and in cases improving in-domain performance.

Double-Hard Debias: Tailoring Word Embeddings for Gender Bias Mitigation [Website][PDF]
Tianlu Wang, Xi Victoria Lin, Nazneen Fatema Rajani, Bryan McCann, Vicente Ordonez, and Caiming Xiong 21:00–22:00

Word embeddings derived from human-generated corpora inherit strong gender bias which can be further amplified by downstream models. Some commonly adopted debiasing approaches, including the seminal Hard Debias algorithm, apply post-processing procedures that project pre-trained word embeddings into a subspace orthogonal to an inferred gender subspace. We discover that semantic-agnostic corpus regularities such as word frequency captured by the word embeddings negatively impact the performance of these algorithms. We propose a simple but effective technique, Double Hard Debias, which purifies the word embeddings against such corpus regularities prior to inferring and removing the gender subspace. Experiments on three bias mitigation benchmarks show that our approach preserves the distributional semantics of the pre-trained word embeddings while reducing gender bias to a significantly larger degree than prior approaches.

Gender Bias in Multilingual Embeddings and Cross-Lingual Transfer [Website][PDF]
Jieyu Zhao, Subhabrata Mukherjee, saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah 21:00–22:00

Multilingual representations embed words from many languages into a single semantic space such that words with similar meanings are close to each other regardless of the language. These embeddings have been widely used in various settings, such as cross-lingual transfer, where a natural language processing (NLP) model trained on one language is deployed to another language. While the cross-lingual transfer techniques are powerful, they carry gender bias from the source to target languages. In this paper, we study gender bias in multilingual embeddings and how it affects transfer learning for NLP applications. We create a multilingual dataset for bias analysis and propose several ways for quantifying bias in multilingual representations from both the intrinsic and extrinsic perspectives. Experimental results show that the magnitude of bias in the multilingual representations changes differently when we align the embeddings to different target spaces and that the alignment direction can also have an influence on the bias in transfer learning. We further provide recommendations for using the multilingual word representations for downstream tasks.

Language (Technology) is Power: A Critical Survey of "Bias" in NLP [Website][PDF]
Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach 21:00–22:00

We survey 146 papers analyzing "bias" in NLP systems, finding that their motivations are often vague, inconsistent, and lacking in normative reasoning, despite the fact that analyzing "bias" is an inherently normative process. We further find that these papers' proposed quantitative techniques for measuring or mitigating "bias" are poorly matched to their motivations and do not engage with the relevant literature outside of NLP. Based on these findings, we describe the beginnings of a path forward by proposing three recommendations that should guide work analyzing "bias" in NLP systems. These recommendations rest on a greater recognition of the relationships between language and social hierarchies, encouraging researchers and practitioners to articulate their conceptualizations of "bias"—i.e., what kinds of system behaviors are harmful, in what ways, to whom, and why, as well as the normative reasoning underlying these statements—and to center work around the lived experiences of members of communities affected by NLP systems, while interrogating and reimagining the power relations between technologists and such communities.

Mitigating Gender Bias Amplification in Distribution by Posterior Regularization [Website][PDF]
Shengyu Jia, Tao Meng, Jieyu Zhao, and Kai-Wei Chang 21:00–22:00

Advanced machine learning techniques have boosted the performance of natural language processing. Nevertheless, recent studies, e.g., (CITATION) show that these techniques inadvertently capture the societal bias hidden in the corpus and further amplify it. However, their analysis is conducted only on models' top predictions. In this paper, we investigate the gender bias amplification issue from the distribution perspective and demonstrate that the bias is amplified in the view of predicted probability distribution over labels. We further propose a bias mitigation approach based on posterior regularization. With little performance loss, our method can almost remove the bias amplification in the distribution. Our study sheds the light on understanding the bias amplification.

Social Bias Frames: Reasoning about Social and Power Implications of Language [Website][PDF]
Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi 21:00–22:00

Warning: this paper contains content that may be offensive or upsetting. Language has the power to reinforce stereotypes and project social biases onto others. At the core of the challenge is that it is rarely what is stated explicitly, but rather the implied meanings, that frame people's judgments about others. For example, given a statement that "we shouldn't lower our standards to hire more women," most listeners will infer the implicature intended by the

speaker - that "women (candidates) are less qualified." Most semantic formalisms, to date, do not capture such pragmatic implications in which people express social biases and power differentials in language. We introduce Social Bias Frames, a new conceptual formalism that aims to model the pragmatic frames in which people project social biases and stereotypes onto others. In addition, we introduce the Social Bias Inference Corpus to support large-scale modelling and evaluation with 150k structured annotations of social media posts, covering over 34k implications about a thousand demographic groups. We then establish baseline approaches that learn to recover Social Bias Frames from unstructured text. We find that while state-of-the-art neural models are effective at high-level categorization of whether a given statement projects unwanted social bias (80% F1), they are not effective at spelling out more detailed explanations in terms of Social Bias Frames. Our study motivates future work that combines structured pragmatic inference with commonsense reasoning on social implications.

Social Biases in NLP Models as Barriers for Persons with Disabilities [Website] [PDF]
Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl 21:00–22:00

Building equitable and inclusive NLP technologies demands consideration of whether and how social attitudes are represented in ML models. In particular, representations encoded in models often inadvertently perpetuate undesirable social biases from the data on which they are trained. In this paper, we present evidence of such undesirable biases towards mentions of disability in two different English language models: toxicity prediction and sentiment analysis. Next, we demonstrate that the neural embeddings that are the critical first step in most NLP pipelines similarly contain undesirable biases towards mentions of disability. We end by highlighting topical biases in the discourse about disability which may contribute to the observed model biases; for instance, gun violence, homelessness, and drug addiction are over-represented in texts discussing mental illness.

Toward Gender-Inclusive Coreference Resolution [Website] [PDF]
Yang Trista Cao and Hal Daumé III 21:00–22:00

Correctly resolving textual mentions of people fundamentally entails making inferences about those people. Such inferences raise the risk of systemic biases in coreference resolution systems, including biases that can harm binary and non-binary trans and cis stakeholders. To better understand such biases, we foreground nuanced conceptualizations of gender from sociology and sociolinguistics, and develop two new datasets for interrogating bias in crowd annotations and in existing coreference resolution systems. Through these studies, conducted on English text, we confirm that without acknowledging and building systems that recognize the complexity of gender, we build systems that lead to many potential harms.

Towards Debiasing Sentence Representations [Website] [PDF]
Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency 21:00–22:00

As natural language processing methods are increasingly deployed in real-world scenarios such as healthcare, legal systems, and social science, it becomes necessary to recognize the role they potentially play in shaping social biases and stereotypes. Previous work has revealed the presence of social biases in widely used word embeddings involving gender, race, religion, and other social constructs. While some methods were proposed to debias these word-level embeddings, there is a need to perform debiasing at the sentence-level given the recent shift towards new contextualized sentence representations such as ELMo and BERT. In this paper, we investigate the presence of social biases in sentence-level representations and propose a new method, Sent-Debias, to reduce these biases. We show that Sent-Debias is effective in removing biases, and at the same time, preserves performance on sentence-level downstream tasks such as sentiment analysis, linguistic acceptability, and natural language understanding. We hope that our work will inspire future research on characterizing and removing social biases from widely adopted sentence representations for fairer NLP.

Session 10B: Interpretability and Analysis of Models for NLP-8

Explaining Black Box Predictions and Unveiling Data Artifacts through Influence Functions [Website][PDF]
Xiaochuang Han, Byron C. Wallace, and Yulia Tsvetkov 21:00–22:00

Modern deep learning models for NLP are notoriously opaque. This has motivated the development of methods for interpreting such models, e.g., via gradient-based saliency maps or the visualization of attention weights. Such approaches aim to provide explanations for a particular model prediction by highlighting important words in the corresponding input text. While this might be useful for tasks where decisions are explicitly influenced by individual tokens in the input, we suspect that such highlighting is not suitable for tasks where model decisions should be driven by more complex reasoning. In this work, we investigate the use of influence functions for NLP, providing an alternative approach to interpreting neural text classifiers. Influence functions explain the decisions of a model by identifying influential training examples. Despite the promise of this approach, influence functions have not yet been extensively evaluated in the context of NLP, a gap addressed by this work. We conduct a comparison between influence functions and common word-saliency methods on representative tasks. As suspected, we find that influence functions are particularly useful for natural language inference, a task in which 'saliency maps' may not have clear interpretation. Furthermore, we develop a new quantitative measure based on influence functions that can reveal artifacts in training data.

Generating Hierarchical Explanations on Text Classification via Feature Interaction Detection [Website][PDF]
Hanjie Chen, Guangtao Zheng, and Yangfeng Ji 21:00–22:00

Generating explanations for neural networks has become crucial for their applications in real-world with respect to reliability and trustworthiness. In natural language processing, existing methods usually provide important features which are words or phrases selected from an input text as an explanation, but ignore the interactions between them. It poses challenges for humans to interpret an explanation and connect it to model prediction. In this work, we build hierarchical explanations by detecting feature interactions. Such explanations visualize how words and phrases are combined at different levels of the hierarchy, which can help users understand the decision-making of black-box models. The proposed method is evaluated with three neural text classifiers (LSTM, CNN, and BERT) on two benchmark datasets, via both automatic and human evaluations. Experiments show the effectiveness of the proposed method in providing explanations that are both faithful to models and interpretable to humans.

Learning to Deceive with Attention-Based Explanations [Website][PDF]
Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C. Lipton 21:00–22:00

Attention mechanisms are ubiquitous components in neural architectures applied to natural language processing. In addition to yielding gains in predictive accuracy, attention weights are often claimed to confer interpretability, purportedly useful both for providing insights to practitioners and for explaining why a model makes its decisions to stakeholders. We call the latter use of attention mechanisms into question by demonstrating a simple method for training models to produce deceptive attention masks. Our method diminishes the total weight assigned to designated impermissible tokens, even when the models can be shown to nevertheless rely on these features to drive predictions. Across multiple models and tasks, our approach manipulates attention weights while paying surprisingly little cost in accuracy. Through a human study, we show that our manipulated attention-based explanations deceive people into thinking that predictions from a model biased against gender minorities do not rely on the gender. Consequently, our results cast doubt on attention's reliability as a tool for auditing algorithms in the context of fairness and accountability.

Learning to Faithfully Rationalize by Construction [Website][PDF]
Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C. Wallace 21:00–22:00

In many settings it is important for one to be able to understand why a model made a particular prediction. In NLP this often entails extracting snippets of an input text 'responsible for' corresponding model output; when such a snippet comprises tokens that indeed informed the model's prediction, it is a faithful explanation. In some settings, faithfulness may be critical to ensure transparency. Lei et al. (2016) proposed a model to produce faithful rationales for neural text classification by defining independent snippet extraction and prediction modules. However, the discrete selection over input tokens performed by this method complicates training, leading to high variance and requiring careful hyperparameter tuning. We propose a simpler variant of this approach that provides faithful explanations by construction. In our scheme, named FRESH, arbitrary feature importance scores (e.g., gradients from a trained model) are used to induce binary labels over token inputs, which an extractor can be trained to predict. An independent classifier module is then trained exclusively on snippets provided by the extractor; these snippets thus constitute faithful explanations, even if the classifier is arbitrarily complex. In both automatic and manual evaluations we find that variants of this simple framework yield predictive performance superior to 'end-to-end' approaches, while being more general and easier to train. Code is available at https://github.com/successar/FRESH.

On the Spontaneous Emergence of Discrete and Compositional Signals [Website][PDF]
Nur Geffen Lan, Emmanuel Chemla, and Shane Steinert-Threlkeld 21:00–22:00

We propose a general framework to study language emergence through signaling games with neural agents. Using a continuous latent space, we are able to (i) train using backpropagation, (ii) show that discrete messages nonetheless naturally emerge. We explore whether categorical perception effects follow and show that the messages are not compositional.

Session 10B: Language Grounding to Vision, Robotics and Beyond-4

Dense-Caption Matching and Frame-Selection Gating for Temporal Localization in VideoQA [Website][PDF]
Hyounghun Kim, Zineng Tang, and Mohit Bansal 21:00–22:00

Videos convey rich information. Dynamic spatio-temporal relationships between people/objects, and diverse multimodal events are present in a video clip. Hence, it is important to develop automated models that can accurately extract such information from videos. Answering questions on videos is one of the tasks which can evaluate such AI abilities. In this paper, we propose a video question answering model which effectively integrates multi-modal input sources and finds the temporally relevant information to answer questions. Specifically, we first employ dense image captions to help identify objects and their detailed salient regions and actions, and hence give the model useful extra information (in explicit textual format to allow easier matching) for answering questions. Moreover, our model is also comprised of dual-level attention (word/object and frame level), multi-head self/cross-integration for different sources (video and dense captions), and gates which pass more relevant information to the classifier. Finally, we also cast the frame selection problem as a multi-label classification task and introduce two loss functions, In-andOut Frame Score Margin (IOFSM) and Balanced Binary Cross-Entropy (BBCE), to better supervise the model with human importance annotations. We evaluate our model on the challenging TVQA dataset, where each of our model components provides significant gains, and our overall model outperforms the state-of-the-art by a large margin (74.09% versus 70.52%). We also present several word, object, and frame level visualization studies.

Shaping Visual Representations with Language for Few-Shot Classification [Website][PDF]
Jesse Mu, Percy Liang, and Noah Goodman 21:00–22:00

By describing the features and abstractions of our world, language is a crucial tool for human learning and a promising source of supervision for machine learning models. We use language to improve few-shot visual classification in the underexplored scenario where natural language task descriptions are available during training, but unavailable for novel tasks at test time. Existing models for this setting sample new descriptions at test time and use those to classify images. Instead, we propose language-shaped learning (LSL), an end-to-end model that regularizes visual representations to predict language. LSL is conceptually simpler, more data efficient, and outperforms baselines in two challenging few-shot domains.

Session 10B: Machine Learning for NLP-11

Discrete Latent Variable Representations for Low-Resource Text Classification [Website][PDF]
Shuning Jin, Sam Wiseman, Karl Stratos, and Karen Livescu 21:00–22:00

While much work on deep latent variable models of text uses continuous latent variables, discrete latent variables are interesting because they are more interpretable and typically more space efficient. We consider several approaches to learning discrete latent variable models for text in the case where exact marginalization over these variables is intractable. We compare the performance of the learned representations as features for low-resource document and sentence classification. Our best models outperform the previous best reported results with continuous representations in these low-resource settings, while learning significantly more compressed representations. Interestingly, we find that an amortized variant of Hard EM performs particularly well in the lowest-resource regimes.

Improving Transformer Models by Reordering their Sublayers [Website][PDF]
Ofir Press, Noah A. Smith, and Omer Levy 21:00–22:00

Multilayer transformer networks consist of interleaved self-attention and feedforward sublayers. Could ordering the sublayers in a different pattern lead to better performance? We generate randomly ordered transformers and train them with the language modeling objective. We observe that some of these models are able to achieve better performance than the interleaved baseline, and that those successful variants tend to have more self-attention at the bottom and more feedforward sublayers at the top. We propose a new transformer pattern that adheres to this property, the sandwich transformer, and show that it improves perplexity on multiple word-level and character-level language modeling benchmarks, at no cost in parameters, memory, or training time. However, the sandwich reordering pattern does not guarantee performance gains across every task, as we demonstrate on machine translation models. Instead, we suggest that further exploration of task-specific sublayer reorderings is needed in order to unlock additional gains.

Learning Constraints for Structured Prediction Using Rectifier Networks [Website][PDF]
Xingyuan Pan, Maitrey Mehta, and Vivek Srikumar 21:00–22:00

Various natural language processing tasks are structured prediction problems where outputs are constructed with multiple interdependent decisions. Past work has shown that domain knowledge, framed as constraints over the output space, can help improve predictive accuracy. However, designing good constraints often relies on domain expertise. In this paper, we study the problem of learning such constraints. We frame the problem as that of training a two-layer rectifier network to identify valid structures or substructures, and show a construction for converting a trained network into a system of linear constraints over the inference variables. Our experiments on several NLP tasks show that the learned constraints can improve the prediction accuracy, especially when the number of training examples is small.

Pretraining with Contrastive Sentence Objectives Improves Discourse Performance of Language Models [Website][PDF]
Dan Iter, Kelvin Guu, Larry Lansing, and Dan Jurafsky 21:00–22:00

Recent models for unsupervised representation learning of text have employed a number of techniques to improve contextual word representations but have put little focus on discourse-level representations. We propose Conpono, an inter-sentence objective for pretraining language models that models discourse coherence and the distance between sentences. Given an anchor sentence, our model is trained to predict the text k sentences away using a sampled-softmax objective where the candidates consist of neighboring sentences and sentences randomly sampled from the corpus. On the discourse representation benchmark DiscoEval, our model improves over the previous state-of-the-art by up to 13% and on average 4% absolute across 7 tasks. Our model is the same size as BERT-Base, but outperforms the much larger BERT-Large model and other more recent approaches that incorporate discourse. We also show that Conpono yields gains of 2%-6% absolute even for tasks that do not explicitly evaluate discourse: textual entailment (RTE), common sense reasoning (COPA) and reading comprehension (ReCoRD).

SAFER: A Structure-free Approach for Certified Robustness to Adversarial Word Substitutions [Website][PDF]
Mao Ye, Chengyue Gong, and Qiang Liu 21:00–22:00

State-of-the-art NLP models can often be fooled by human-unaware transformations such as synonymous word substitution. For security reasons, it is of critical importance to develop models with certified robustness that can provably guarantee that the prediction is can not be altered by any possible synonymous word substitution. In this work, we propose a certified robust method based on a new randomized smoothing technique, which constructs a stochastic ensemble by applying random word substitutions on the input sentences, and leverage the statistical properties of the ensemble to provably certify the robustness. Our method is simple and structure-free in that it only requires the black-box queries of the model outputs, and hence can be applied to any pre-trained models (such as BERT) and any types of models (world-level or subword-level). Our method significantly outperforms recent state-of-the-art methods for certified robustness on both IMDB and Amazon text classification tasks. To the best of our knowledge, we are the first work to achieve certified robustness on large systems such as BERT with practically meaningful certified accuracy.

[TACL] SpanBERT: Improving Pre-training by Representing and Predicting Spans [Website][PDF]
Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy 21:00–22:00

We present SpanBERT, a pre-training method that is designed to better represent and predict spans of text. Our approach extends BERT by (1) masking contiguous random spans, rather than random tokens, and (2) training the span

boundary representations to predict the entire content of the masked span, without relying on the individual token representations within it. SpanBERT consistently outperforms BERT and our better-tuned baselines, with substantial gains on span selection tasks such as question answering and coreference resolution. In particular, with the same training data and model size as BERT-Large, our single model obtains 94.6% and 88.7% F1 on SQuAD 1.1 and 2.0 respectively. We also achieve a new state of the art on the OntoNotes coreference resolution task (79.6% F1), strong performance on the TACRED relation extraction benchmark, and even gains on GLUE.

Session 10B: Question Answering-8

Clinical Reading Comprehension: A Thorough Analysis of the emrQA Dataset [Website][PDF]
Xiang Yue, Bernal Jimenez Gutierrez, and Huan Sun 21:00–22:00

Machine reading comprehension has made great progress in recent years owing to large-scale annotated datasets. In the clinical domain, however, creating such datasets is quite difficult due to the domain expertise required for annotation. Recently, Pampari et al. (EMNLP'18) tackled this issue by using expert-annotated question templates and existing i2b2 annotations to create emrQA, the first large-scale dataset for question answering (QA) based on clinical notes. In this paper, we provide an in-depth analysis of this dataset and the clinical reading comprehension (CliniRC) task. From our qualitative analysis, we find that (i) emrQA answers are often incomplete, and (ii) emrQA questions are often answerable without using domain knowledge. From our quantitative experiments, surprising results include that (iii) using a small sampled subset (5%-20%), we can obtain roughly equal performance compared to the model trained on the entire dataset, (iv) this performance is close to human expert's performance, and (v) BERT models do not beat the best performing base model. Following our analysis of the emrQA, we further explore two desired aspects of CliniRC systems: the ability to utilize clinical domain knowledge and to generalize to unseen questions and contexts. We argue that both should be considered when creating future datasets.

On the Importance of Diversity in Question Generation for QA [Website][PDF]
Md Arafat Sultan, Shubham Chandel, Ramón Fernandez Astudillo, and Vittorio Castelli 21:00–22:00

Automatic question generation (QG) has shown promise as a source of synthetic training data for question answering (QA). In this paper we ask: Is textual diversity in QG beneficial for downstream QA? Using top-p nucleus sampling to derive samples from a transformer-based question generator, we show that diversity-promoting QG indeed provides better QA training than likelihood maximization approaches such as beam search. We also show that standard QG evaluation metrics such as BLEU, ROUGE and METEOR are inversely correlated with diversity, and propose a diversity-aware intrinsic measure of overall QG quality that correlates well with extrinsic evaluation on QA.

SCDE: Sentence Cloze Dataset with High Quality Distractors From Examinations [Website][PDF]
Xiang Kong, Varun Gangal, and Eduard Hovy 21:00–22:00

We introduce SCDE, a dataset to evaluate the performance of computational models through sentence prediction. SCDE is a human created sentence cloze dataset, collected from public school English examinations. Our task requires a model to fill up multiple blanks in a passage from a shared candidate set with distractors designed by English teachers. Experimental results demonstrate that this task requires the use of non-local, discourse-level context beyond the immediate sentence neighborhood. The blanks require joint solving and significantly impair each other's context. Furthermore, through ablations, we show that the distractors are of high quality and make the task more challenging. Our experiments show that there is a significant performance gap between advanced models (72%) and humans (87%), encouraging future models to bridge this gap.

The Cascade Transformer: an Application for Efficient Answer Sentence Selection [Website][PDF]
Luca Soldaini and Alessandro Moschitti 21:00–22:00

Large transformer-based language models have been shown to be very effective in many classification tasks. However, their computational complexity prevents their use in applications requiring the classification of a large set of candidates. While previous works have investigated approaches to reduce model size, relatively little attention has been paid to techniques to improve batch throughput during inference. In this paper, we introduce the Cascade Transformer, a simple yet effective technique to adapt transformer-based models into a cascade of rankers. Each ranker is used to prune a subset of candidates in a batch, thus dramatically increasing throughput at inference time. Partial encodings from the transformer model are shared among rerankers, providing further speed-up. When compared to a state-of-the-art transformer model, our approach reduces computation by 37% with almost no impact on accuracy, as measured on two English Question Answering datasets.

Transformers to Learn Hierarchical Contexts in Multiparty Dialogue for Span-based Question Answering [Website][PDF]
Changmao Li and Jinho D. Choi 21:00–22:00

We introduce a novel approach to transformers that learns hierarchical representations in multiparty dialogue. First, three language modeling tasks are used to pre-train the transformers, token- and utterance-level language modeling and utterance order prediction, that learn both token and utterance embeddings for better understanding in dialogue contexts. Then, multi-task learning between the utterance prediction and the token span prediction is applied to fine-tune for span-based question answering (QA). Our approach is evaluated on the FriendsQA dataset and shows improvements of 3.8% and 1.4% over the two state-of-the-art transformer models, BERT and RoBERTa, respectively.

Session 10B: Resources and Evaluation-12

A Recipe for Creating Multimodal Aligned Datasets for Sequential Tasks [Website][PDF]
Angela Lin, Sudha Rao, Asli Celikyilmaz, Elnaz Nouri, Chris Brockett, Debadeepta Dey, and Bill Dolan
21:00–22:00

Many high-level procedural tasks can be decomposed into sequences of instructions that vary in their order and choice of tools. In the cooking domain, the web offers many, partially-overlapping, text and video recipes (i.e. procedures) that describe how to make the same dish (i.e. high-level task). Aligning instructions for the same dish across different sources can yield descriptive visual explanations that are far richer semantically than conventional textual instructions, providing commonsense insight into how real-world procedures are structured. Learning to align these different instruction sets is challenging because: a) different recipes vary in their order of instructions and use of ingredients; and b) video instructions can be noisy and tend to contain far more information than text instructions. To address these challenges, we use an unsupervised alignment algorithm that learns pairwise alignments between instructions of different recipes for the same dish. We then use a graph algorithm to derive a joint alignment between multiple text and multiple video recipes for the same dish. We release the MICROSOFT RESEARCH MULTIMODAL ALIGNED RECIPE CORPUS containing ~150K pairwise alignments between recipes across 4262 dishes with rich commonsense information.

Adversarial NLI: A New Benchmark for Natural Language Understanding [Website][PDF]
Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela 21:00–22:00

We introduce a new large-scale NLI benchmark dataset, collected via an iterative, adversarial human-and-model-in-the-loop procedure. We show that training models on this new dataset leads to state-of-the-art performance on a variety of popular NLI benchmarks, while posing a more difficult challenge with its new test set. Our analysis sheds light on the shortcomings of current state-of-the-art models, and shows that non-expert annotators are successful at finding their weaknesses. The data collection method can be applied in a never-ending learning scenario, becoming a moving target for NLU, rather than a static benchmark that will quickly saturate.

Dialogue-Based Relation Extraction [Website][PDF]
Dian Yu, Kai Sun, Claire Cardie, and Dong Yu 21:00–22:00

We present the first human-annotated dialogue-based relation extraction (RE) dataset DialogRE, aiming to support the prediction of relation(s) between two arguments that appear in a dialogue. We further offer DialogRE as a platform for studying cross-sentence RE as most facts span multiple sentences. We argue that speaker-related information plays a critical role in the proposed task, based on an analysis of similarities and differences between dialogue-based and traditional RE tasks. Considering the timeliness of communication in a dialogue, we design a new metric to evaluate the performance of RE methods in a conversational setting and investigate the performance of several representative RE methods on DialogRE. Experimental results demonstrate that a speaker-aware extension on the best-performing model leads to gains in both the standard and conversational evaluation settings. DialogRE is available at https://dataset.org/dialogre/.

Dscorer: A Fast Evaluation Metric for Discourse Representation Structure Parsing [Website][PDF]
Jiangming Liu, Shay B. Cohen, and Mirella Lapata 21:00–22:00

Discourse representation structures (DRSs) are scoped semantic representations for texts of arbitrary length. Evaluating the accuracy of predicted DRSs plays a key role in developing semantic parsers and improving their performance. DRSs are typically visualized as boxes which are not straightforward to process automatically. Counter transforms DRSs to clauses and measures clause overlap by searching for variable mappings between two DRSs. However, this metric is computationally costly (with respect to memory and CPU time) and does not scale with longer texts. We introduce Dscorer, an efficient new metric which converts box-style DRSs to graphs and then measures the overlap of n-grams. Experiments show that Dscorer computes accuracy scores that are correlated with Counter at a fraction of the time.

Facet-Aware Evaluation for Extractive Summarization [Website][PDF]
Yuning Mao, Liyuan Liu, Qi Zhu, Xiang Ren, and Jiawei Han 21:00–22:00

Commonly adopted metrics for extractive summarization focus on lexical overlap at the token level. In this paper, we present a facet-aware evaluation setup for better assessment of the information coverage in extracted summaries. Specifically, we treat each sentence in the reference summary as a *facet*, identify the sentences in the document that express the semantics of each facet as *support sentences* of the facet, and automatically evaluate extractive summarization methods by comparing the indices of extracted sentences and support sentences of all the facets in the reference summary. To facilitate this new evaluation setup, we construct an extractive version of the CNN/Daily Mail dataset and perform a thorough quantitative investigation, through which we demonstrate that facet-aware evaluation manifests better correlation with human judgment than ROUGE, enables fine-grained evaluation as well as comparative analysis, and reveals valuable insights of state-of-the-art summarization methods. Data can be found at `https://github.com/morningmoni/FAR`.

More Diverse Dialogue Datasets via Diversity-Informed Data Collection [Website][PDF]
Katherine Stasaski, Grace Hui Yang, and Marti A. Hearst 21:00–22:00

Automated generation of conversational dialogue using modern neural architectures has made notable advances. However, these models are known to have a drawback of often producing uninteresting, predictable responses; this is known as the diversity problem. We introduce a new strategy to address this problem, called Diversity-Informed Data Collection. Unlike prior approaches, which modify model architectures to solve the problem, this method uses

dynamically computed corpus-level statistics to determine which conversational participants to collect data from. Diversity-Informed Data Collection produces significantly more diverse data than baseline data collection methods, and better results on two downstream tasks: emotion classification and dialogue generation. This method is generalizable and can be used with other corpus-level metrics.

Not All Claims are Created Equal: Choosing the Right Statistical Approach to Assess Hypotheses [Website][PDF]
Erfan Sadeqi Azer, Daniel Khashabi, Ashish Sabharwal, and Dan Roth 21:00–22:00

Empirical research in Natural Language Processing (NLP) has adopted a narrow set of principles for assessing hypotheses, relying mainly on p-value computation, which suffers from several known issues. While alternative proposals have been well-debated and adopted in other fields, they remain rarely discussed or used within the NLP community. We address this gap by contrasting various hypothesis assessment techniques, especially those not commonly used in the field (such as evaluations based on Bayesian inference). Since these statistical techniques differ in the hypotheses they can support, we argue that practitioners should first decide their target hypothesis before choosing an assessment method. This is crucial because common fallacies, misconceptions, and misinterpretation surrounding hypothesis assessment methods often stem from a discrepancy between what one would like to claim versus what the method used actually assesses. Our survey reveals that these issues are omnipresent in the NLP research community. As a step forward, we provide best practices and guidelines tailored to NLP research, as well as an easy-to-use package for Bayesian assessment of hypotheses, complementing existing tools.

S2ORC: The Semantic Scholar Open Research Corpus [Website][PDF]
Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld 21:00–22:00

We introduce S2ORC, a large corpus of 81.1M English-language academic papers spanning many academic disciplines. The corpus consists of rich metadata, paper abstracts, resolved bibliographic references, as well as structured full text for 8.1M open access papers. Full text is annotated with automatically-detected inline mentions of citations, figures, and tables, each linked to their corresponding paper objects. In S2ORC, we aggregate papers from hundreds of academic publishers and digital archives into a unified source, and create the largest publicly-available collection of machine-readable academic text to date. We hope this resource will facilitate research and development of tools and tasks for text mining over academic text.

STARC: Structured Annotations for Reading Comprehension [Website][PDF]
Yevgeni Berzak, Jonathan Malmaud, and Roger Levy 21:00–22:00

We present STARC (Structured Annotations for Reading Comprehension), a new annotation framework for assessing reading comprehension with multiple choice questions. Our framework introduces a principled structure for the answer choices and ties them to textual span annotations. The framework is implemented in OneStopQA, a new high-quality dataset for evaluation and analysis of reading comprehension in English. We use this dataset to demonstrate that STARC can be leveraged for a key new application for the development of SAT-like reading comprehension materials: automatic annotation quality probing via span ablation experiments. We further show that it enables in-depth analyses and comparisons between machine and human reading comprehension behavior, including error distributions and guessing ability. Our experiments also reveal that the standard multiple choice dataset in NLP, RACE, is limited in its ability to measure reading comprehension. 47% of its questions can be guessed by machines without accessing the passage, and 18% are unanimously judged by humans as not having a unique correct answer. OneStopQA provides an alternative test set for reading comprehension which alleviates these shortcomings and has a substantially higher human ceiling performance.

WinoWhy: A Deep Diagnosis of Essential Commonsense Knowledge for Answering Winograd Schema Challenge [Website][PDF]
Hongming Zhang, Xinran Zhao, and Yangqiu Song 21:00–22:00

In this paper, we present the first comprehensive categorization of essential commonsense knowledge for answering the Winograd Schema Challenge (WSC). For each of the questions, we invite annotators to first provide reasons for making correct decisions and then categorize them into six major knowledge categories. By doing so, we better understand the limitation of existing methods (i.e., what kind of knowledge cannot be effectively represented or inferred with existing methods) and shed some light on the commonsense knowledge that we need to acquire in the future for better commonsense reasoning. Moreover, to investigate whether current WSC models can understand the commonsense or they simply solve the WSC questions based on the statistical bias of the dataset, we leverage the collected reasons to develop a new task called WinoWhy, which requires models to distinguish plausible reasons from very similar but wrong reasons for all WSC questions. Experimental results prove that even though pre-trained language representation models have achieved promising progress on the original WSC dataset, they are still struggling at WinoWhy. Further experiments show that even though supervised models can achieve better performance, the performance of these models can be sensitive to the dataset distribution. WinoWhy and all codes are available at: https://github.com/HKUST-KnowComp/WinoWhy.

Session 10B: Speech and Multimodality-6

Towards end-2-end learning for predicting behavior codes from spoken utterances in psychotherapy conversations [Website][PDF]
Karan Singla, Zhuohao Chen, David Atkins, and Shrikanth Narayanan 21:00–22:00

Spoken language understanding tasks usually rely on pipelines involving complex processing blocks such as voice activity detection, speaker diarization and Automatic speech recognition (ASR). We propose a novel framework for predicting utterance level labels directly from speech features, thus removing the dependency on first generating transcripts, and transcription free behavioral coding. Our classifier uses a pretrained Speech-2-Vector encoder as bottleneck to generate word-level representations from speech features. This pretrained encoder learns to encode speech features for a word using an objective similar to Word2Vec. Our proposed approach just uses speech features and word segmentation information for predicting spoken utterance-level target labels. We show that our model achieves competitive results to other state-of-the-art approaches which use transcribed text for the task of predicting psychotherapy-relevant behavior codes.

Session 10B: Summarization-5

A Transformer-based Approach for Source Code Summarization [Website][PDF]
Wasi Ahmad, Saikat Chakraborty, Baishakhi Ray, and Kai-Wei Chang 21:00–22:00

Generating a readable summary that describes the functionality of a program is known as source code summarization. In this task, learning code representation by modeling the pairwise relationship between code tokens to capture their long-range dependencies is crucial. To learn code representation for summarization, we explore the Transformer model that uses a self-attention mechanism and has shown to be effective in capturing long-range dependencies. In this work, we show that despite the approach is simple, it outperforms the state-of-the-art techniques by a significant margin. We perform extensive analysis and ablation studies that reveal several important findings, e.g., the absolute encoding of source code tokens' position hinders, while relative encoding significantly improves the summarization performance. We have made our code publicly available[4] to facilitate future research.

Asking and Answering Questions to Evaluate the Factual Consistency of Summaries [Website][PDF]
Alex Wang, Kyunghyun Cho, and Mike Lewis 21:00–22:00

Practical applications of abstractive summarization models are limited by frequent factual inconsistencies with respect to their input. Existing automatic evaluation metrics for summarization are largely insensitive to such errors. We propose QAGS (pronounced "kags"), an automatic evaluation protocol that is designed to identify factual inconsistencies in a generated summary. QAGS is based on the intuition that if we ask questions about a summary and its source, we will receive similar answers if the summary is factually consistent with the source. To evaluate QAGS, we collect human judgments of factual consistency on model-generated summaries for the CNN/DailyMail (Hermann et al., 2015) and XSUM (Narayan et al., 2018) summarization datasets. QAGS has substantially higher correlations with these judgments than other automatic evaluation metrics. Also, QAGS offers a natural form of interpretability: The answers and questions generated while computing QAGS indicate which tokens of a summary are inconsistent and why. We believe QAGS is a promising tool in automatically generating usable and factually consistent text. Code for QAGS will be available at https://github.com/W4ngatang/qags.

Discourse-Aware Neural Extractive Text Summarization [Website][PDF]
Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu 21:00–22:00

Recently BERT has been adopted for document encoding in state-of-the-art text summarization models. However, sentence-based extractive models often result in redundant or uninformative phrases in the extracted summaries. Also, long-range dependencies throughout a document are not well captured by BERT, which is pre-trained on sentence pairs instead of documents. To address these issues, we present a discourse-aware neural summarization model - DiscoBert. DiscoBert extracts sub-sentential discourse units (instead of sentences) as candidates for extractive selection on a finer granularity. To capture the long-range dependencies among discourse units, structural discourse graphs are constructed based on RST trees and coreference mentions, encoded with Graph Convolutional Networks. Experiments show that the proposed model outperforms state-of-the-art methods by a significant margin on popular summarization benchmarks compared to other BERT-base models.

Discrete Optimization for Unsupervised Sentence Summarization with Word-Level Extraction [Website][PDF]
Raphael Schumann, Lili Mou, Yao Lu, Olga Vechtomova, and Katja Markert 21:00–22:00

Automatic sentence summarization produces a shorter version of a sentence, while preserving its most important information. A good summary is characterized by language fluency and high information overlap with the source sentence. We model these two aspects in an unsupervised objective function, consisting of language modeling and semantic similarity metrics. We search for a high-scoring summary by discrete optimization. Our proposed method achieves a new state-of-the art for unsupervised sentence summarization according to ROUGE scores. Additionally, we demonstrate that the commonly reported ROUGE F1 metric is sensitive to summary length. Since this is unwillingly exploited in recent work, we emphasize that future evaluation should explicitly group summarization systems by output length brackets.

Exploring Content Selection in Summarization of Novel Chapters [Website][PDF]
Faisal Ladhak, Bryan Li, Yaser Al-Onaizan, and Kathleen McKeown 21:00–22:00

We present a new summarization task, generating summaries of novel chapters using summary/chapter pairs from online study guides. This is a harder task than the news summarization task, given the chapter length as well as the extreme paraphrasing and generalization found in the summaries. We focus on extractive summarization, which requires the creation of a gold-standard set of extractive summaries. We present a new metric for aligning reference summary sentences with chapter sentences to create gold extracts and also experiment with different alignment methods. Our experiments demonstrate significant improvement over prior alignment approaches for our task as shown through automatic metrics and a crowd-sourced pyramid analysis.

FEQA: A Question Answering Evaluation Framework for Faithfulness Assessment in Abstractive Summarization [Website][PDF]
Esin Durmus, He He, and Mona Diab 21:00–22:00

Neural abstractive summarization models are prone to generate content inconsistent with the source document, i.e. unfaithful. Existing automatic metrics do not capture such mistakes effectively. We tackle the problem of evaluating

[4] https://github.com/wasiahmad/NeuralCodeSum

faithfulness of a generated summary given its source document. We first collected human annotations of faithfulness for outputs from numerous models on two datasets. We find that current models exhibit a trade-off between abstractiveness and faithfulness: outputs with less word overlap with the source document are more likely to be unfaithful. Next, we propose an automatic question answering (QA) based metric for faithfulness, FEQA, which leverages recent advances in reading comprehension. Given question-answer pairs generated from the summary, a QA model extracts answers from the document; non-matched answers indicate unfaithful information in the summary. Among metrics based on word overlap, embedding similarity, and learned language understanding models, our QA-based metric has significantly higher correlation with human faithfulness scores, especially on highly abstractive summaries.

Fact-based Content Weighting for Evaluating Abstractive Summarisation [Website][PDF]
Xinnuo Xu, Ondřej Dušek, Jingyi Li, Verena Rieser, and Ioannis Konstas 21:00–22:00

Abstractive summarisation is notoriously hard to evaluate since standard word-overlap-based metrics are insufficient. We introduce a new evaluation metric which is based on fact-level content weighting, i.e. relating the facts of the document to the facts of the summary. We fol- low the assumption that a good summary will reflect all relevant facts, i.e. the ones present in the ground truth (human-generated refer- ence summary). We confirm this hypothe- sis by showing that our weightings are highly correlated to human perception and compare favourably to the recent manual highlight- based metric of Hardy et al. (2019).

Hooks in the Headline: Learning to Generate Headlines with Controlled Styles [Website][PDF]
Di Jin, Zhijing Jin, Joey Tianyi Zhou, Lisa Orii, and Peter Szolovits 21:00–22:00

Current summarization systems only produce plain, factual headlines, far from the practical needs for the exposure and memorableness of the articles. We propose a new task, Stylistic Headline Generation (SHG), to enrich the headlines with three style options (humor, romance and clickbait), thus attracting more readers. With no style-specific article-headline pair (only a standard headline summarization dataset and mono-style corpora), our method TitleStylist generates stylistic headlines by combining the summarization and reconstruction tasks into a multitasking framework. We also introduced a novel parameter sharing scheme to further disentangle the style from text. Through both automatic and human evaluation, we demonstrate that TitleStylist can generate relevant, fluent headlines with three target styles: humor, romance, and clickbait. The attraction score of our model generated headlines outperforms the state-of-the-art summarization model by 9.68%, even outperforming human-written references.

Knowledge Graph-Augmented Abstractive Summarization with Semantic-Driven Cloze Reward [Website][PDF]
Luyang Huang, Lingfei Wu, and Lu Wang 21:00–22:00

Sequence-to-sequence models for abstractive summarization have been studied extensively, yet the generated summaries commonly suffer from fabricated content, and are often found to be near-extractive. We argue that, to address these issues, the summarizer should acquire semantic interpretation over input, e.g., via structured representation, to allow the generation of more informative summaries. In this paper, we present ASGARD, a novel framework for Abstractive Summarization with Graph-Augmentation and semantic-driven RewarD. We propose the use of dual encoders—a sequential document encoder and a graph-structured encoder—to maintain the global context and local characteristics of entities, complementing each other. We further design a reward based on a multiple choice cloze test to drive the model to better capture entity interactions. Results show that our models produce significantly higher ROUGE scores than a variant without knowledge graph as input on both New York Times and CNN/Daily Mail datasets. We also obtain better or comparable performance compared to systems that are fine-tuned from large pretrained language models. Human judges further rate our model outputs as more informative and containing fewer unfaithful errors.

Optimizing the Factual Correctness of a Summary: A Study of Summarizing Radiology Reports [Website][PDF]
Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D. Manning, and Curtis Langlotz 21:00–22:00

Neural abstractive summarization models are able to generate summaries which have high overlap with human references. However, existing models are not optimized for factual correctness, a critical metric in real-world applications. In this work, we develop a general framework where we evaluate the factual correctness of a generated summary by fact-checking it automatically against its reference using an information extraction module. We further propose a training strategy which optimizes a neural summarization model with a factual correctness reward via reinforcement learning. We apply the proposed method to the summarization of radiology reports, where factual correctness is a key requirement. On two separate datasets collected from hospitals, we show via both automatic and human evaluation that the proposed approach substantially improves the factual correctness and overall quality of outputs over a competitive neural summarization system, producing radiology summaries that approach the quality of human-authored ones.

Storytelling with Dialogue: A Critical Role Dungeons and Dragons Dataset [Website][PDF]
Revanth Rameshkumar and Peter Bailey 21:00–22:00

This paper describes the Critical Role Dungeons and Dragons Dataset (CRD3) and related analyses. Critical Role is an unscripted, live-streamed show where a fixed group of people play Dungeons and Dragons, an open-ended role-playing game. The dataset is collected from 159 Critical Role episodes transcribed to text dialogues, consisting of 398,682 turns. It also includes corresponding abstractive summaries collected from the Fandom wiki. The dataset is linguistically unique in that the narratives are generated entirely through player collaboration and spoken interaction. For each dialogue, there are a large number of turns, multiple abstractive summaries with varying levels of detail, and semantic ties to the previous dialogues. In addition, we provide a data augmentation method that produces 34,243 summary-dialogue chunk pairs to support current neural ML approaches, and we provide an abstractive summariza-

tion benchmark and evaluation.

The Summary Loop: Learning to Write Abstractive Summaries Without Examples [Website][PDF]
Philippe Laban, Andrew Hsi, John Canny, and Marti A. Hearst 21:00–22:00

This work presents a new approach to unsupervised abstractive summarization based on maximizing a combination of coverage and fluency for a given length constraint. It introduces a novel method that encourages the inclusion of key terms from the original document into the summary: key terms are masked out of the original document and must be filled in by a coverage model using the current generated summary. A novel unsupervised training procedure leverages this coverage model along with a fluency model to generate and score summaries. When tested on popular news summarization datasets, the method outperforms previous unsupervised methods by more than 2 R-1 points, and approaches results of competitive supervised methods. Our model attains higher levels of abstraction with copied passages roughly two times shorter than prior work, and learns to compress and merge sentences without supervision.

Unsupervised Opinion Summarization as Copycat-Review Generation [Website][PDF]
Arthur Bražinskas, Mirella Lapata, and Ivan Titov 21:00–22:00

Opinion summarization is the task of automatically creating summaries that reflect subjective information expressed in multiple documents, such as product reviews. While the majority of previous work has focused on the extractive setting, i.e., selecting fragments from input reviews to produce a summary, we let the model generate novel sentences and hence produce abstractive summaries. Recent progress in summarization has seen the development of supervised models which rely on large quantities of document-summary pairs. Since such training data is expensive to acquire, we instead consider the unsupervised setting, in other words, we do not use any summaries in training. We define a generative model for a review collection which capitalizes on the intuition that when generating a new review given a set of other reviews of a product, we should be able to control the "amount of novelty" going into the new review or, equivalently, vary the extent to which it deviates from the input. At test time, when generating summaries, we force the novelty to be minimal, and produce a text reflecting consensus opinions. We capture this intuition by defining a hierarchical variational autoencoder model. Both individual reviews and the products they correspond to are associated with stochastic latent codes, and the review generator ("decoder") has direct access to the text of input reviews through the pointer-generator mechanism. Experiments on Amazon and Yelp datasets, show that setting at test time the review's latent code to its mean, allows the model to produce fluent and coherent summaries reflecting common opinions.

Demo Session 5C

Time: *21:30–22:15*

NLP Scholar: An Interactive Visual Explorer for Natural Language Processing Literature [Website][PDF]
Saif M. Mohammad

As part of the NLP Scholar project, we created a single unified dataset of NLP papers and their meta-information (including citation numbers), by extracting and aligning information from the ACL Anthology and Google Scholar. In this paper, we describe several interconnected interactive visualizations (dashboards) that present various aspects of the data. Clicking on an item within a visualization or entering query terms in the search boxes filters the data in all visualizations in the dashboard. This allows users to search for papers in the area of their interest, published within specific time periods, published by specified authors, etc. The interactive visualizations presented here, and the associated dataset of papers mapped to citations, have additional uses as well including understanding how the field is growing (both overall and across sub-areas), as well as quantifying the impact of different types of papers on subsequent publications.

Stimulating Creativity with FunLines: A Case Study of Humor Generation in Headlines [Website][PDF]
Nabil Hossain, John Krumm, Tanvir Sajed, and Henry Kautz

Building datasets of creative text, such as humor, is quite challenging. We introduce FunLines, a competitive game where players edit news headlines to make them funny, and where they rate the funniness of headlines edited by others. FunLines makes the humor generation process fun, interactive, collaborative, rewarding and educational, keeping players engaged and providing humor data at a very low cost compared to traditional crowdsourcing approaches. FunLines offers useful performance feedback, assisting players in getting better over time at generating and assessing humor, as our analysis shows. This helps to further increase the quality of the generated dataset. We show the effectiveness of this data by training humor classification models that outperform a previous benchmark, and we release this dataset to the public.

Usnea: An Authorship Tool for Interactive Fiction using Retrieval Based Semantic Parsing [Website][PDF]
Ben Swanson and Boris Smus

The reader of a choose your own adventure novel and the user of a modern virtual assistant have a subtle similarity; both may, through the right lens, be viewed as engaging with a work of Interactive Fiction. This literary form emerged in the 1970s and has grown like a vine along the branch of modern technology, one guided by the advances of the other. In this work we weave together threads from the Interactive Fiction community and neural semantic parsing for dialog systems, defining the data model and necessary algorithms for a novel type of Interactive Fiction and open sourcing its accompanying authoring tool. Specifically, our work integrates retrieval based semantic parsing predicates into the branching story structures well known to the Interactive Fiction community, relaxing the relatively strict lexical options of preexisting systems.

5

Main Conference: Wednesday, July 8

Overview

5:00–5:45 **Demo Session 1A**

5:00–6:00 **Session 11A**
Dialogue and Interactive Systems-13
Information Extraction-3
Machine Translation-13
NLP Applications-8
Sentence Level-5
Textual Inference and Other Areas of Semantics-3
Student Research Workshop
Summarization-6
Tagging, Chunking and Parsing-4
Theme-3

5:45–6:30 **Demo Session 1B**

6:00–7:00 **Session 11B**
Dialogue and Interactive Systems-14
Discourse and Pragmatics-6
Information Extraction-4
Language Grounding to Vision, Robotics and Beyond-5
Machine Learning for NLP-12
Phonology, Morphology and Word Segmentation-3
Question Answering-9
Sentence Level-6
Student Research Workshop

6:30–7:15 **Demo Session 1C**

8:00–8:45 **Demo Session 2A**

8:00–9:00 **Session 12A**
Discourse and Pragmatics-7
Information Extraction-5
Information Retrieval and Text Mining-6
Machine Learning for NLP-13
Machine Translation-14
NLP Applications-9
Sentence Level-7
Sentiment Analysis, Stylistic Analysis, and Argument Mining-10
Student Research Workshop
Summarization-7

8:45–9:30 **Demo Session 2B**

9:00–10:00 **Session 12B**
Dialogue and Interactive Systems-15
Generation-11
Information Extraction-6
Language Grounding to Vision, Robotics and Beyond-6
Machine Learning for NLP-14
Phonology, Morphology and Word Segmentation-4
Question Answering-10
Textual Inference and Other Areas of Semantics-4
Student Research Workshop
Theme-4

9:30–10:15 **Demo Session 2C**
12:00–12:45 **Demo Session 3A**
12:00–13:00 **Session 13A**
Generation-12
Information Extraction-7
Machine Learning for NLP-15
NLP Applications-10
Lexical-7
Sentence Level-8
Textual Inference and Other Areas of Semantics-5
Sentiment Analysis, Stylistic Analysis, and Argument Mining-11
Student Research Workshop

12:45–13:30 **Demo Session 3B**
13:00–14:00 **Session 13B**
Dialogue and Interactive Systems-16
Discourse and Pragmatics-8
Information Extraction-8
Language Grounding to Vision, Robotics and Beyond-7
Machine Translation-15
Phonology, Morphology and Word Segmentation-5
Question Answering-11
Student Research Workshop
Theme-5

13:30–14:15 **Demo Session 3C**
14:00–14:45 **Keynote 2 Video Livestream: Josh Tenenbaum (Sponsored by DeepMind and Google)**
14:45–15:15 **Keynote 2 Live Q&A: Josh Tenenbaum (Sponsored by DeepMind and Google)**
15:15–15:25 **Best Paper Award Ceremony**
15:25–15:37 **Future Conferences**
15:37–15:49 **Closing Remarks**
17:00–17:45 **Demo Session 4A**
17:00–18:00 **Session 14A**
Generation-13
Information Extraction-9
Information Retrieval and Text Mining-7
Language Grounding to Vision, Robotics and Beyond-8
Phonology, Morphology and Word Segmentation-6
Sentence Level-9
Student Research Workshop

17:45–18:30 **Demo Session 4B**
18:00–19:00 **Session 14B**
Information Extraction-10
Machine Learning for NLP-16
Machine Translation-16
NLP Applications-11
Lexical-8
Textual Inference and Other Areas of Semantics-6
Student Research Workshop
Tagging, Chunking and Parsing-5

18:30–19:15 **Demo Session 4C**
20:00–20:45 **Demo Session 5A**

20:00–21:00 **Session 15A**
Generation-14
Information Extraction-11
Information Retrieval and Text Mining-8
Language Grounding to Vision, Robotics and Beyond-9
Machine Translation-17
Sentence Level-10
Textual Inference and Other Areas of Semantics-7
Student Research Workshop

20:45–21:30 **Demo Session 5B**

21:00–22:00 **Session 15B**
Generation-15
Information Extraction-12
Machine Translation-18
NLP Applications-12
Phonology, Morphology and Word Segmentation-7
Sentence Level-11
Student Research Workshop
Tagging, Chunking and Parsing-6
Theme-6

21:30–22:15 **Demo Session 5C**

Demo Session 1A

Time: *5:00–5:45*

ADVISER: A Toolkit for Developing Multi-modal, Multi-domain and Socially-engaged Conversational Agents [Website] [PDF]
Chia-Yu Li, Daniel Ortega, Dirk Väth, Florian Lux, Lindsey Vanderlyn, Maximilian Schmidt, Michael Neumann, Moritz Völkel, Pavel Denisov, Sabrina Jenne, Zorica Kacarevic, and Ngoc Thang Vu

We present ADVISER - an open-source, multi-domain dialog system toolkit that enables the development of multi-modal (incorporating speech, text and vision), socially-engaged (e.g. emotion recognition, engagement level prediction and backchanneling) conversational agents. The final Python-based implementation of our toolkit is flexible, easy to use, and easy to extend not only for technically experienced users, such as machine learning researchers, but also for less technically experienced users, such as linguists or cognitive scientists, thereby providing a flexible platform for collaborative research.

Prta: A System to Support the Analysis of Propaganda Techniques in the News [Website] [PDF]
Giovanni Da San Martino, Shaden Shaar, Yifan Zhang, Seunghak Yu, Alberto Barrón-Cedeño, and Preslav Nakov

Recent events, such as the 2016 US Presidential Campaign, Brexit and the COVID-19 "infodemic", have brought into the spotlight the dangers of online disinformation. There has been a lot of research focusing on fact-checking and disinformation detection. However, little attention has been paid to the specific rhetorical and psychological techniques used to convey propaganda messages. Revealing the use of such techniques can help promote media literacy and critical thinking, and eventually contribute to limiting the impact of "fake news" and disinformation campaigns. Prta (Propaganda Persuasion Techniques Analyzer) allows users to explore the articles crawled on a regular basis by highlighting the spans in which propaganda techniques occur and to compare them on the basis of their use of propaganda techniques. The system further reports statistics about the use of such techniques, overall and over time, or according to filtering criteria specified by the user based on time interval, keywords, and/or political orientation of the media. Moreover, it allows users to analyze any text or URL through a dedicated interface or via an API. The system is available online: https://www.tanbih.org/prta.

Session 11A Overview – Wednesday, July 8, 2020 5:00–6:00

Track A *Dialogue and Interactive Systems-13* Abstracts	Diverse and Informative Dialogue Generation with Context-Specific Commonsense Knowledge Awareness *Wu, Li, Zhang, Zhou, and Wu* [Website][PDF]	Generate, Delete and Rewrite: A Three-Stage Framework for Improving Persona Consistency of Dialogue Generation *Song, Wang, Zhang, Liu, and Liu* [Website][PDF]	Learning to Customize Model Structures for Few-shot Dialogue Generation Tasks *SONG, Liu, Bi, Yan, and Zhang* [Website][PDF]	Video-Grounded Dialogues with Pretrained Generation Language Models *Le and Hoi* [Website][PDF]	
Track B *Information Extraction-3* Abstracts	A Unified MRC Framework for Named Entity Recognition *Li, Feng, Meng, Han, Wu, and Li* [Website][PDF]	An Effective Transition-based Model for Discontinuous NER *Dai, Karimi, Hachey, and Paris* [Website][PDF]	IMoJIE: Iterative Memory-Based Joint Open Information Extraction *Kolluru, Aggarwal, Rathore, and Chakrabarti* [Website][PDF]	Improving Event Detection via Open-domain Trigger Knowledge *Tong, Xu, Wang, Cao, Hou, Li, and Xie* [Website][PDF]	Improving Low-Resource Named Entity Recognition using Joint Sentence and Token Labeling *Kruengkrai, Nguyen, Aljunied, and Bing* [Website][PDF]
	Multi-Cell Compositional LSTM for NER Domain Adaptation *Jia and Zhang* [Website][PDF]	Pyramid: A Layered Model for Nested Named Entity Recognition *WANG, Shou, Chen, and Chen* [Website][PDF]	ReInceptionE: Relation-Aware Inception Network with Joint Local-Global Structural Information for Knowledge Graph Embedding *Xie, Zhou, Liu, and Huang* [Website][PDF]	Relabel the Noise: Joint Extraction of Entities and Relations via Cooperative Multiagents *Chen, Li, Lei, and Shen* [Website][PDF]	Simplify the Usage of Lexicon in Chinese NER *Ma, Peng, Zhang, Wei, and Huang* [Website][PDF]
Track C *Machine Translation-13* Abstracts	AdvAug: Robust Adversarial Augmentation for Neural Machine Translation *Cheng, Jiang, Macherey, and Eisenstein* [Website][PDF]	Contextual Neural Machine Translation Improves Translation of Cataphoric Pronouns *Wong, Maruf, and Haffari* [Website][PDF]	Improving Neural Machine Translation with Soft Template Prediction *Yang, Ma, Zhang, Li, and Zhou* [Website][PDF]	Tagged Back-translation Revisited: Why Does It Really Work? *Marie, Rubino, and Fujita* [Website][PDF]	Worse WER, but Better BLEU? Leveraging Word Embedding as Intermediate in Multitask End-to-End Speech Translation *Chuang, Sung, Liu, and Lee* [Website][PDF]
Track D *NLP Applications-8* Abstracts	Neural-DINF: A Neural Network based Framework for Measuring Document Influence *Tan, Yang, Li, Tang, Huang, and Zhuang* [Website][PDF]	Paraphrase Generation by Learning How to Edit from Samples *Kazemnejad, Salehi, and Soleymani Baghshah* [Website][PDF]			

Track E *Sentence Level-5* Abstracts	Emerging Cross-lingual Structure in Pretrained Language Models *Conneau, Wu, Li, Zettlemoyer, and Stoyanov* [Website][PDF]	FastBERT: a Self-distilling BERT with Adaptive Inference Time *Liu, Zhou, Wang, Zhao, Deng, and JU* [Website][PDF]	Incorporating External Knowledge through Pre-training for Natural Language to Code Generation *Xu, Jiang, Yin, Vasilescu, and Neubig* [Website][PDF]	LogicalFactChecker: Leveraging Logical Operations for Fact Checking with Graph Module Network *Zhong, Tang, Feng, Duan, Zhou, Gong, Shou, Jiang, Wang, and Yin* [Website][PDF]	Word-level Textual Adversarial Attacking as Combinatorial Optimization *Zang, Qi, Yang, Liu, Zhang, Liu, and Sun* [Website][PDF]
Track F *Textual Inference and Other Areas of Semantics-3* Abstracts	Benchmarking Multimodal Regex Synthesis with Complex Structures *Ye, Chen, Dillig, and Durrett* [Website][PDF]	Curriculum Learning for Natural Language Understanding *Xu, Zhang, Mao, Wang, Xie, and Zhang* [Website][PDF]	Do Neural Models Learn Systematicity of Monotonicity Inference in Natural Language? *Yanaka, Mineshima, Bekki, and Inui* [Website][PDF]	Evidence-Aware Inferential Text Generation with Vector Quantised Variational AutoEncoder *Guo, Tang, Duan, Yin, Jiang, and Zhou* [Website][PDF]	How to Ask Good Questions? Try to Leverage Paraphrases *Jia, Zhou, SUN, and Wu* [Website][PDF]
	NeuInfer: Knowledge Inference on N-ary Facts *Guan, Jin, Guo, Wang, and Cheng* [Website][PDF]	Neural Graph Matching Networks for Chinese Short Text Matching *Chen, Zhao, Lyu, Jin, Chen, Zhu, and Yu* [Website][PDF]	Neural Mixed Counting Models for Dispersed Topic Discovery *Wu, Rao, Zhang, Xie, Li, Wang, and Chen* [Website][PDF]	Reasoning Over Semantic-Level Graph for Fact Checking *Zhong, Xu, Tang, Xu, Duan, Zhou, Wang, and Yin* [Website][PDF]	
Track G *Student Research Workshop* Abstracts	HGCN4MeSH: Hybrid Graph Convolution Network for MeSH Indexing *Yu, Yang, and Li* [Website][PDF]	Considering Likelihood in NLP Classification Explanations with Occlusion and Language Modeling *Harbecke and Alt* [Website][PDF]			
Track H *Summarization-6* Abstracts	Automatic Generation of Citation Texts in Scholarly Papers: A Pilot Study *Xing, Fan, and Wan* [Website][PDF]	Composing Elementary Discourse Units in Abstractive Summarization *Li, Wu, and Li* [Website][PDF]	Extractive Summarization as Text Matching *Zhong, Liu, Chen, Wang, Qiu, and Huang* [Website][PDF]	Heterogeneous Graph Neural Networks for Extractive Document Summarization *Wang, Liu, Zheng, Qiu, and Huang* [Website][PDF]	Jointly Learning to Align and Summarize for Neural Cross-Lingual Summarization *Cao, Liu, and Wan* [Website][PDF]
	Leveraging Graph to Improve Abstractive Multi-Document Summarization *Li, Xiao, Liu, Wu, Wang, and Du* [Website][PDF]	Multi-Granularity Interaction Network for Extractive and Abstractive Multi-Document Summarization *Jin, Wang, and Wan* [Website][PDF]			
Track I *Tagging, Chunking and Parsing-4* Abstracts	Tetra-Tagging: Word-Synchronous Parsing with Linear-Time Inference *Kitaev and Klein* [Website][PDF]				

Track J *Theme-3* Abstracts	Are we Estimating or Guesstimating Translation Quality? *Sun, Guzmán, and Specia* [Website][PDF]	Language (Re)modelling: Towards Embodied Language Understanding *Tamari, Shani, Hope, Petruck, Abend, and Shahaf* [Website][PDF]	The State and Fate of Linguistic Diversity and Inclusion in the NLP World *Joshi, Santy, Budhiraja, Bali, and Choudhury* [Website][PDF]	The Unstoppable Rise of Computational Linguistics in Deep Learning *Henderson* [Website][PDF]	To Boldly Query What No One Has Annotated Before? The Frontiers of Corpus Querying *Gärtner and Jung* [Website][PDF]

Session 11A Details

Session 11A: Dialogue and Interactive Systems-13

Diverse and Informative Dialogue Generation with Context-Specific Commonsense Knowledge Awareness [Website][PDF]
Sixing Wu, Ying Li, Dawei Zhang, Yang Zhou, and Zhonghai Wu 5:00–6:00

Generative dialogue systems tend to produce generic responses, which often leads to boring conversations. For alleviating this issue, Recent studies proposed to retrieve and introduce knowledge facts from knowledge graphs. While this paradigm works to a certain extent, it usually retrieves knowledge facts only based on the entity word itself, without considering the specific dialogue context. Thus, the introduction of the context-irrelevant knowledge facts can impact the quality of generations. To this end, this paper proposes a novel commonsense knowledge-aware dialogue generation model, ConKADI. We design a Felicitous Fact mechanism to help the model focus on the knowledge facts that are highly relevant to the context; furthermore, two techniques, Context-Knowledge Fusion and Flexible Mode Fusion are proposed to facilitate the integration of the knowledge in the ConKADI. We collect and build a large-scale Chinese dataset aligned with the commonsense knowledge for dialogue generation. Extensive evaluations over both an open-released English dataset and our Chinese dataset demonstrate that our approach ConKADI outperforms the state-of-the-art approach CCM, in most experiments.

Generate, Delete and Rewrite: A Three-Stage Framework for Improving Persona Consistency of Dialogue Generation [Website][PDF]
Haoyu Song, Yan Wang, Wei-Nan Zhang, Xiaojiang Liu, and Ting Liu 5:00–6:00

Maintaining a consistent personality in conversations is quite natural for human beings, but is still a non-trivial task for machines. The persona-based dialogue generation task is thus introduced to tackle the personality-inconsistent problem by incorporating explicit persona text into dialogue generation models. Despite the success of existing persona-based models on generating human-like responses, their one-stage decoding framework can hardly avoid the generation of inconsistent persona words. In this work, we introduce a three-stage framework that employs a generate-delete-rewrite mechanism to delete inconsistent words from a generated response prototype and further rewrite it to a personality-consistent one. We carry out evaluations by both human and automatic metrics. Experiments on the Persona-Chat dataset show that our approach achieves good performance.

Learning to Customize Model Structures for Few-shot Dialogue Generation Tasks [Website][PDF]
YIPING SONG, Zequn Liu, Wei Bi, Rui Yan, and Ming Zhang 5:00–6:00

Training the generative models with minimal corpus is one of the critical challenges for building open-domain dialogue systems. Existing methods tend to use the meta-learning framework which pre-trains the parameters on all non-target tasks then fine-tunes on the target task. However, fine-tuning distinguishes tasks from the parameter perspective but ignores the model-structure perspective, resulting in similar dialogue models for different tasks. In this paper, we propose an algorithm that can customize a unique dialogue model for each task in the few-shot setting. In our approach, each dialogue model consists of a shared module, a gating module, and a private module. The first two modules are shared among all the tasks, while the third one will differentiate into different network structures to better capture the characteristics of the corresponding task. The extensive experiments on two datasets show that our method outperforms all the baselines in terms of task consistency, response quality, and diversity.

Video-Grounded Dialogues with Pretrained Generation Language Models [Website][PDF]
Hung Le and Steven C.H. Hoi 5:00–6:00

Pre-trained language models have shown remarkable success in improving various downstream NLP tasks due to their ability to capture dependencies in textual data and generate natural responses. In this paper, we leverage the power of pre-trained language models for improving video-grounded dialogue, which is very challenging and involves complex features of different dynamics: (1) Video features which can extend across both spatial and temporal dimensions; and (2) Dialogue features which involve semantic dependencies over multiple dialogue turns. We propose a framework by extending GPT-2 models to tackle these challenges by formulating video-grounded dialogue tasks as a sequence-to-sequence task, combining both visual and textual representation into a structured sequence, and fine-tuning a large pre-trained GPT-2 network. Our framework allows fine-tuning language models to capture dependencies across multiple modalities over different levels of information: spatio-temporal level in video and token-sentence level in dialogue context. We achieve promising improvement on the Audio-Visual Scene-Aware Dialogues (AVSD) benchmark from DSTC7, which supports a potential direction in this line of research.

Session 11A: Information Extraction-3

A Unified MRC Framework for Named Entity Recognition [Website][PDF]
Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li 5:00–6:00

The task of named entity recognition (NER) is normally divided into nested NER and flat NER depending on whether named entities are nested or not.Models are usually separately developed for the two tasks, since sequence labeling models, the most widely used backbone for flat NER, are only able to assign a single label to a particular token, which is unsuitable for nested NER where a token may be assigned several labels. In this paper, we propose a unified framework that is capable of handling both flat and nested NER tasks. Instead of treating the task of NER as a sequence labeling problem, we propose to formulate it as a machine reading comprehension (MRC) task. For example, extracting entities with the PER label is formalized as extracting answer spans to the question "*which person is mentioned in the text*".This formulation naturally tackles the entity overlapping issue in nested NER: the extraction of two overlapping entities with different categories requires answering two independent questions. Additionally, since the query encodes informative prior knowledge, this strategy facilitates the process of entity extraction, leading to better performances for not only nested NER, but flat NER. We conduct experiments on both nested and flat NER datasets.Experiment results demonstrate the effectiveness of the proposed formulation. We are able to achieve a vast amount of performance boost over current SOTA models on nested NER datasets, i.e., +1.28, +2.55, +5.44, +6.37,respectively on ACE04, ACE05, GENIA and KBP17, along with SOTA results on flat NER datasets, i.e., +0.24, +1.95, +0.21, +1.49 respectively on English CoNLL 2003, English OntoNotes 5.0, Chinese MSRA and Chinese OntoNotes 4.0.

An Effective Transition-based Model for Discontinuous NER [Website][PDF]
Xiang Dai, Sarvnaz Karimi, Ben Hachey, and Cecile Paris 5:00–6:00

Unlike widely used Named Entity Recognition (NER) data sets in generic domains, biomedical NER data sets often contain mentions consisting of discontinuous spans. Conventional sequence tagging techniques encode Markov assumptions that are efficient but preclude recovery of these mentions. We propose a simple, effective transition-based model with generic neural encoding for discontinuous NER. Through extensive experiments on three biomedical data sets, we show that our model can effectively recognize discontinuous mentions without sacrificing the accuracy on continuous mentions.

IMoJIE: Iterative Memory-Based Joint Open Information Extraction [Website][PDF]
Keshav Kolluru, Samarth Aggarwal, Vipul Rathore, and Soumen Chakrabarti 5:00–6:00

While traditional systems for Open Information Extraction were statistical and rule-based, recently neural models have been introduced for the task. Our work builds upon CopyAttention, a sequence generation OpenIE model (Cui et. al. 18). Our analysis reveals that CopyAttention produces a constant number of extractions per sentence, and its extracted tuples often express redundant information. We present IMoJIE, an extension to CopyAttention, which produces the next extraction conditioned on all previously extracted tuples. This approach overcomes both shortcomings of CopyAttention, resulting in a variable number of diverse extractions per sentence. We train IMoJIE on training data bootstrapped from extractions of several non-neural systems, which have been automatically filtered to reduce redundancy and noise. IMoJIE outperforms CopyAttention by about 18 F1 pts, and a BERT-based strong baseline by 2 F1 pts, establishing a new state of the art for the task.

Improving Event Detection via Open-domain Trigger Knowledge [Website][PDF]
Meihan Tong, Bin Xu, Shuai Wang, Yixin Cao, Lei Hou, Juanzi Li, and Jun Xie 5:00–6:00

Event Detection (ED) is a fundamental task in automatically structuring texts. Due to the small scale of training data, previous methods perform poorly on unseen/sparsely labeled trigger words and are prone to overfitting densely labeled trigger words. To address the issue, we propose a novel Enrichment Knowledge Distillation (EKD) model to leverage external open-domain trigger knowledge to reduce the in-built biases to frequent trigger words in annotations. Experiments on benchmark ACE2005 show that our model outperforms nine strong baselines, is especially effective for unseen/sparsely labeled trigger words. The source code is released on https://github.com/shuaiwa16/ekd.git.

Improving Low-Resource Named Entity Recognition using Joint Sentence and Token Labeling [Website][PDF]
Canasai Kruengkrai, Thien Hai Nguyen, Sharifah Mahani Aljunied, and Lidong Bing 5:00–6:00

Exploiting sentence-level labels, which are easy to obtain, is one of the plausible methods to improve low-resource named entity recognition (NER), where token-level labels are costly to annotate. Current models for jointly learning sentence and token labeling are limited to binary classification. We present a joint model that supports multi-class classification and introduce a simple variant of self-attention that allows the model to learn scaling factors. Our model produces 3.78%, 4.20%, 2.08% improvements in F1 over the BiLSTM-CRF baseline on e-commerce product titles in three different low-resource languages: Vietnamese, Thai, and Indonesian, respectively.

Multi-Cell Compositional LSTM for NER Domain Adaptation [Website][PDF]
Chen Jia and Yue Zhang 5:00–6:00

Cross-domain NER is a challenging yet practical problem. Entity mentions can be highly different across domains. However, the correlations between entity types can be relatively more stable across domains. We investigate a multi-cell compositional LSTM structure for multi-task learning, modeling each entity type using a separate cell state. With the help of entity typed units, cross-domain knowledge transfer can be made in an entity type level. Theoretically, the resulting distinct feature distributions for each entity type make it more powerful for cross-domain transfer. Empirically, experiments on four few-shot and zero-shot datasets show our method significantly outperforms a series of

multi-task learning methods and achieves the best results.

Pyramid: A Layered Model for Nested Named Entity Recognition [Website][PDF]
Jue WANG, Lidan Shou, Ke Chen, and Gang Chen 5:00–6:00

This paper presents Pyramid, a novel layered model for Nested Named Entity Recognition (nested NER). In our approach, token or text region embeddings are recursively inputted into L flat NER layers, from bottom to top, stacked in a pyramid shape. Each time an embedding passes through a layer of the pyramid, its length is reduced by one. Its hidden state at layer l represents an l-gram in the input text, which is labeled only if its corresponding text region represents a complete entity mention. We also design an inverse pyramid to allow bidirectional interaction between layers. The proposed method achieves state-of-the-art F1 scores in nested NER on ACE-2004, ACE-2005, GENIA, and NNE, which are 80.27, 79.42, 77.78, and 93.70 with conventional embeddings, and 87.74, 86.34, 79.31, and 94.68 with pre-trained contextualized embeddings. In addition, our model can be used for the more general task of Overlapping Named Entity Recognition. A preliminary experiment confirms the effectiveness of our method in overlapping NER.

ReInceptionE: Relation-Aware Inception Network with Joint Local-Global Structural Information for Knowledge Graph Embedding [Website][PDF]
Zhiwen Xie, Guangyou Zhou, Jin Liu, and Jimmy Xiangji Huang 5:00–6:00

The goal of Knowledge graph embedding (KGE) is to learn how to represent the low dimensional vectors for entities and relations based on the observed triples. The conventional shallow models are limited to their expressiveness. ConvE (Dettmers et al., 2018) takes advantage of CNN and improves the expressive power with parameter efficient operators by increasing the interactions between head and relation embeddings. However, there is no structural information in the embedding space of ConvE, and the performance is still limited by the number of interactions. The recent KBGAT (Nathani et al., 2019) provides another way to learn embeddings by adaptively utilizing structural information. In this paper, we take the benefits of ConvE and KBGAT together and propose a Relation-aware Inception network with joint local-global structural information for knowledge graph Embedding (ReInceptionE). Specifically, we first explore the Inception network to learn query embedding, which aims to further increase the interactions between head and relation embeddings. Then, we propose to use a relation-aware attention mechanism to enrich the query embedding with the local neighborhood and global entity information. Experimental results on both WN18RR and FB15k-237 datasets demonstrate that ReInceptionE achieves competitive performance compared with state-of-the-art methods.

Relabel the Noise: Joint Extraction of Entities and Relations via Cooperative Multiagents [Website][PDF]
Daoyuan Chen, Yaliang Li, Kai Lei, and Ying Shen 5:00–6:00

Distant supervision based methods for entity and relation extraction have received increasing popularity due to the fact that these methods require light human annotation efforts. In this paper, we consider the problem of shifted label distribution, which is caused by the inconsistency between the noisy-labeled training set subject to external knowledge graph and the human-annotated test set, and exacerbated by the pipelined entity-then-relation extraction manner with noise propagation. We propose a joint extraction approach to address this problem by re-labeling noisy instances with a group of cooperative multiagents. To handle noisy instances in a fine-grained manner, each agent in the cooperative group evaluates the instance by calculating a continuous confidence score from its own perspective; To leverage the correlations between these two extraction tasks, a confidence consensus module is designed to gather the wisdom of all agents and re-distribute the noisy training set with confidence-scored labels. Further, the confidences are used to adjust the training losses of extractors. Experimental results on two real-world datasets verify the benefits of re-labeling noisy instance, and show that the proposed model significantly outperforms the state-of-the-art entity and relation extraction methods.

Simplify the Usage of Lexicon in Chinese NER [Website][PDF]
Ruotian Ma, Minlong Peng, Qi Zhang, Zhongyu Wei, and Xuanjing Huang 5:00–6:00

Recently, many works have tried to augment the performance of Chinese named entity recognition (NER) using word lexicons. As a representative, Lattice-LSTM has achieved new benchmark results on several public Chinese NER datasets. However, Lattice-LSTM has a complex model architecture. This limits its application in many industrial areas where real-time NER responses are needed. In this work, we propose a simple but effective method for incorporating the word lexicon into the character representations. This method avoids designing a complicated sequence modeling architecture, and for any neural NER model, it requires only subtle adjustment of the character representation layer to introduce the lexicon information. Experimental studies on four benchmark Chinese NER datasets show that our method achieves an inference speed up to 6.15 times faster than those of state-of-the-art methods, along with a better performance. The experimental results also show that the proposed method can be easily incorporated with pre-trained models like BERT.

Session 11A: Machine Translation-13

AdvAug: Robust Adversarial Augmentation for Neural Machine Translation [Website][PDF]
Yong Cheng, Lu Jiang, Wolfgang Macherey, and Jacob Eisenstein 5:00–6:00

In this paper, we propose a new adversarial augmentation method for Neural Machine Translation (NMT). The main idea is to minimize the vicinal risk over virtual sentences sampled from two vicinity distributions, in which the crucial one is a novel vicinity distribution for adversarial sentences that describes a smooth interpolated embedding space centered around observed training sentence pairs. We then discuss our approach, AdvAug, to train NMT models using the embeddings of virtual sentences in sequence-to-sequence learning. Experiments on Chinese-English, English-French, and English-German translation benchmarks show that AdvAug achieves significant improvements over theTransformer (up to 4.9 BLEU points), and substantially outperforms other data augmentation techniques (e.g.back-translation) without using extra corpora.

Contextual Neural Machine Translation Improves Translation of Cataphoric Pronouns [Website][PDF]
KayYen Wong, Sameen Maruf, and Gholamreza Haffari 5:00–6:00

The advent of context-aware NMT has resulted in promising improvements in the overall translation quality and specifically in the translation of discourse phenomena such as pronouns. Previous works have mainly focused on the use of past sentences as context with a focus on anaphora translation. In this work, we investigate the effect of future sentences as context by comparing the performance of a contextual NMT model trained with the future context to the one trained with the past context. Our experiments and evaluation, using generic and pronoun-focused automatic metrics, show that the use of future context not only achieves significant improvements over the context-agnostic Transformer, but also demonstrates comparable and in some cases improved performance over its counterpart trained on past context. We also perform an evaluation on a targeted cataphora test suite and report significant gains over the context-agnostic Transformer in terms of BLEU.

Improving Neural Machine Translation with Soft Template Prediction [Website][PDF]
Jian Yang, Shuming Ma, Dongdong Zhang, Zhoujun Li, and Ming Zhou 5:00–6:00

Although neural machine translation (NMT) has achieved significant progress in recent years, most previous NMT models only depend on the source text to generate translation. Inspired by the success of template-based and syntax-based approaches in other fields, we propose to use extracted templates from tree structures as soft target templates to guide the translation procedure. In order to learn the syntactic structure of the target sentences, we adopt constituency-based parse tree to generate candidate templates. We incorporate the template information into the encoder-decoder framework to jointly utilize the templates and source text. Experiments show that our model significantly outperforms the baseline models on four benchmarks and demonstrates the effectiveness of soft target templates.

Tagged Back-translation Revisited: Why Does It Really Work? [Website][PDF]
Benjamin Marie, Raphael Rubino, and Atsushi Fujita 5:00–6:00

In this paper, we show that neural machine translation (NMT) systems trained on large back-translated data overfit some of the characteristics of machine-translated texts. Such NMT systems better translate human-produced translations, i.e., translationese, but may largely worsen the translation quality of original texts. Our analysis reveals that adding a simple tag to back-translations prevents this quality degradation and improves on average the overall translation quality by helping the NMT system to distinguish back-translated data from original parallel data during training. We also show that, in contrast to high-resource configurations, NMT systems trained in low-resource settings are much less vulnerable to overfit back-translations. We conclude that the back-translations in the training data should always be tagged especially when the origin of the text to be translated is unknown.

Worse WER, but Better BLEU? Leveraging Word Embedding as Intermediate in Multitask End-to-End Speech Translation [Website][PDF]
Shun-Po Chuang, Tzu-Wei Sung, Alexander H. Liu, and Hung-yi Lee 5:00–6:00

Speech translation (ST) aims to learn transformations from speech in the source language to the text in the target language. Previous works show that multitask learning improves the ST performance, in which the recognition decoder generates the text of the source language, and the translation decoder obtains the final translations based on the output of the recognition decoder. Because whether the output of the recognition decoder has the correct semantics is more critical than its accuracy, we propose to improve the multitask ST model by utilizing word embedding as the intermediate.

Session 11A: NLP Applications-8

Neural-DINF: A Neural Network based Framework for Measuring Document Influence [Website] [PDF]
Jie Tan, Changlin Yang, Ying Li, Siliang Tang, Chen Huang, and Yueting Zhuang 5:00–6:00

Measuring the scholarly impact of a document without citations is an important and challenging problem. Existing approaches such as Document Influence Model (DIM) are based on dynamic topic models, which only consider the word frequency change. In this paper, we use both frequency changes and word semantic shifts to measure document influence by developing a neural network framework. Our model has three steps. Firstly, we train the word embeddings for different time periods. Subsequently, we propose an unsupervised method to align vectors for different time periods. Finally, we compute the influence value of documents. Our experimental results show that our model outperforms DIM.

Paraphrase Generation by Learning How to Edit from Samples [Website] [PDF]
Amirhossein Kazemnejad, Mohammadreza Salehi, and Mahdieh Soleymani Baghshah 5:00–6:00

Neural sequence to sequence text generation has been proved to be a viable approach to paraphrase generation. Despite promising results, paraphrases generated by these models mostly suffer from lack of quality and diversity. To address these problems, we propose a novel retrieval-based method for paraphrase generation. Our model first retrieves a paraphrase pair similar to the input sentence from a pre-defined index. With its novel editor module, the model then paraphrases the input sequence by editing it using the extracted relations between the retrieved pair of sentences. In order to have fine-grained control over the editing process, our model uses the newly introduced concept of Micro Edit Vectors. It both extracts and exploits these vectors using the attention mechanism in the Transformer architecture. Experimental results show the superiority of our paraphrase generation method in terms of both automatic metrics, and human evaluation of relevance, grammaticality, and diversity of generated paraphrases.

Session 11A Semantics: Sentence Level-5

Emerging Cross-lingual Structure in Pretrained Language Models [Website][PDF]
Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov 5:00–6:00

We study the problem of multilingual masked language modeling, i.e. the training of a single model on concatenated text from multiple languages, and present a detailed study of several factors that influence why these models are so effective for cross-lingual transfer. We show, contrary to what was previously hypothesized, that transfer is possible even when there is no shared vocabulary across the monolingual corpora and also when the text comes from very different domains. The only requirement is that there are some shared parameters in the top layers of the multilingual encoder. To better understand this result, we also show that representations from monolingual BERT models in different languages can be aligned post-hoc quite effectively, strongly suggesting that, much like for non-contextual word embeddings, there are universal latent symmetries in the learned embedding spaces. For multilingual masked language modeling, these symmetries are automatically discovered and aligned during the joint training process.

FastBERT: a Self-distilling BERT with Adaptive Inference Time [Website][PDF]
Weijie Liu, Peng Zhou, Zhiruo Wang, Zhe Zhao, Haotang Deng, and QI JU 5:00–6:00

Pre-trained language models like BERT have proven to be highly performant. However, they are often computationally expensive in many practical scenarios, for such heavy models can hardly be readily implemented with limited resources. To improve their efficiency with an assured model performance, we propose a novel speed-tunable FastBERT with adaptive inference time. The speed at inference can be flexibly adjusted under varying demands, while redundant calculation of samples is avoided. Moreover, this model adopts a unique self-distillation mechanism at fine-tuning, further enabling a greater computational efficacy with minimal loss in performance. Our model achieves promising results in twelve English and Chinese datasets. It is able to speed up by a wide range from 1 to 12 times than BERT if given different speedup thresholds to make a speed-performance tradeoff.

Incorporating External Knowledge through Pre-training for Natural Language to Code Generation [Website][PDF]
Frank F. Xu, Zhengbao Jiang, Pengcheng Yin, Bogdan Vasilescu, and Graham Neubig 5:00–6:00

Open-domain code generation aims to generate code in a general-purpose programming language (such as Python) from natural language (NL) intents. Motivated by the intuition that developers usually retrieve resources on the web when writing code, we explore the effectiveness of incorporating two varieties of external knowledge into NL-to-code generation: automatically mined NL-code pairs from the online programming QA forum StackOverflow and programming language API documentation. Our evaluations show that combining the two sources with data augmentation and retrieval-based data re-sampling improves the current state-of-the-art by up to 2.2% absolute BLEU score on the code generation testbed CoNaLa. The code and resources are available at https://github.com/neulab/external-knowledge-codegen.

LogicalFactChecker: Leveraging Logical Operations for Fact Checking with Graph Module Network [Website][PDF]
Wanjun Zhong, Duyu Tang, Zhangyin Feng, Nan Duan, Ming Zhou, Ming Gong, Linjun Shou, Daxin Jiang, Jiahai Wang, and Jian Yin 5:00–6:00

Verifying the correctness of a textual statement requires not only semantic reasoning about the meaning of words, but also symbolic reasoning about logical operations like count, superlative, aggregation, etc. In this work, we propose LogicalFactChecker, a neural network approach capable of leveraging logical operations for fact checking. It achieves the state-of-the-art performance on TABFACT, a large-scale, benchmark dataset built for verifying a textual statement with semi-structured tables. This is achieved by a graph module network built upon the Transformer-based architecture. With a textual statement and a table as the input, LogicalFactChecker automatically derives a program (a.k.a. logical form) of the statement in a semantic parsing manner. A heterogeneous graph is then constructed to capture not only the structures of the table and the program, but also the connections between inputs with different modalities. Such a graph reveals the related contexts of each word in the statement, the table and the program. The graph is used to obtain graph-enhanced contextual representations of words in Transformer-based architecture. After that, a program-driven module network is further introduced to exploit the hierarchical structure of the program, where semantic compositionality is dynamically modeled along the program structure with a set of function-specific modules. Ablation experiments suggest that both the heterogeneous graph and the module network are important to obtain strong results.

Word-level Textual Adversarial Attacking as Combinatorial Optimization [Website][PDF]
Yuan Zang, Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Meng Zhang, Qun Liu, and Maosong Sun 5:00–6:00

Adversarial attacks are carried out to reveal the vulnerability of deep neural networks. Textual adversarial attacking is challenging because text is discrete and a small perturbation can bring significant change to the original input. Word-level attacking, which can be regarded as a combinatorial optimization problem, is a well-studied class of textual attack methods. However, existing word-level attack models are far from perfect, largely because unsuitable search space reduction methods and inefficient optimization algorithms are employed. In this paper, we propose a novel attack model, which incorporates the sememe-based word substitution method and particle swarm optimization-based search algorithm to solve the two problems separately. We conduct exhaustive experiments to evaluate our attack model by attacking BiLSTM and BERT on three benchmark datasets. Experimental results demonstrate that our model consistently achieves much higher attack success rates and crafts more high-quality adversarial examples

as compared to baseline methods. Also, further experiments show our model has higher transferability and can bring more robustness enhancement to victim models by adversarial training. All the code and data of this paper can be obtained on https://github.com/thunlp/SememePSO-Attack.

Session 11A Semantics: Textual Inference and Other Areas of Semantics-3

Benchmarking Multimodal Regex Synthesis with Complex Structures [Website][PDF]
Xi Ye, Qiaochu Chen, Isil Dillig, and Greg Durrett 5:00–6:00

Existing datasets for regular expression (regex) generation from natural language are limited in complexity; compared to regex tasks that users post on StackOverflow, the regexes in these datasets are simple, and the language used to describe them is not diverse. We introduce StructuredRegex, a new regex synthesis dataset differing from prior ones in three aspects. First, to obtain structurally complex and realistic regexes, we generate the regexes using a probabilistic grammar with pre-defined macros observed from real-world StackOverflow posts. Second, to obtain linguistically diverse natural language descriptions, we show crowdworkers abstract depictions of the underlying regex and ask them to describe the pattern they see, rather than having them paraphrase synthetic language. Third, we augment each regex example with a collection of strings that are and are not matched by the ground truth regex, similar to how real users give examples. Our quantitative and qualitative analysis demonstrates the advantages of StructuredRegex over prior datasets. Further experimental results using various multimodal synthesis techniques highlight the challenge presented by our dataset, including non-local constraints and multi-modal inputs.

Curriculum Learning for Natural Language Understanding [Website][PDF]
Benfeng Xu, Licheng Zhang, Zhendong Mao, Quan Wang, Hongtao Xie, and Yongdong Zhang 5:00–6:00

With the great success of pre-trained language models, the pretrain-finetune paradigm now becomes the undoubtedly dominant solution for natural language understanding (NLU) tasks. At the fine-tune stage, target task data is usually introduced in a completely random order and treated equally. However, examples in NLU tasks can vary greatly in difficulty, and similar to human learning procedure, language models can benefit from an easy-to-difficult curriculum. Based on this idea, we propose our Curriculum Learning approach. By reviewing the trainset in a crossed way, we are able to distinguish easy examples from difficult ones, and arrange a curriculum for language models. Without any manual model architecture design or use of external data, our Curriculum Learning approach obtains significant and universal performance improvements on a wide range of NLU tasks.

Do Neural Models Learn Systematicity of Monotonicity Inference in Natural Language? [Website][PDF]
Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, and Kentaro Inui 5:00–6:00

Despite the success of language models using neural networks, it remains unclear to what extent neural models have the generalization ability to perform inferences. In this paper, we introduce a method for evaluating whether neural models can learn systematicity of monotonicity inference in natural language, namely, the regularity for performing arbitrary inferences with generalization on composition. We consider four aspects of monotonicity inferences and test whether the models can systematically interpret lexical and logical phenomena on different training/test splits. A series of experiments show that three neural models systematically draw inferences on unseen combinations of lexical and logical phenomena when the syntactic structures of the sentences are similar between the training and test sets. However, the performance of the models significantly decreases when the structures are slightly changed in the test set while retaining all vocabularies and constituents already appearing in the training set. This indicates that the generalization ability of neural models is limited to cases where the syntactic structures are nearly the same as those in the training set.

Evidence-Aware Inferential Text Generation with Vector Quantised Variational AutoEncoder [Website][PDF]
Daya Guo, Duyu Tang, Nan Duan, Jian Yin, Daxin Jiang, and Ming Zhou 5:00–6:00

Generating inferential texts about an event in different perspectives requires reasoning over different contexts that the event occurs. Existing works usually ignore the context that is not explicitly provided, resulting in a context-independent semantic representation that struggles to support the generation. To address this, we propose an approach that automatically finds evidence for an event from a large text corpus, and leverages the evidence to guide the generation of inferential texts. Our approach works in an encoderdecoder manner and is equipped with Vector Quantised-Variational Autoencoder, where the encoder outputs representations from a distribution over discrete variables. Such discrete representations enable automatically selecting relevant evidence, which not only facilitates evidence-aware generation, but also provides a natural way to uncover rationales behind the generation. Our approach provides state-of-the-art performance on both Event2mind and Atomic datasets. More importantly, we find that with discrete representations, our model selectively uses evidence to generate different inferential texts.

How to Ask Good Questions? Try to Leverage Paraphrases [Website][PDF]
Xin Jia, Wenjie Zhou, Xu SUN, and Yunfang Wu 5:00–6:00

Given a sentence and its relevant answer, how to ask good questions is a challenging task, which has many real applications. Inspired by human's paraphrasing capability to ask questions of the same meaning but with diverse expressions, we propose to incorporate paraphrase knowledge into question generation(QG) to generate human-like questions. Specifically, we present a two-hand hybrid model leveraging a self-built paraphrase resource, which is automatically conducted by a simple back-translation method. On the one hand, we conduct multi-task learning with sentence-level paraphrase generation (PG) as an auxiliary task to supplement paraphrase knowledge to the task-share encoder. On the other hand, we adopt a new loss function for diversity training to introduce more question patterns to QG. Extensive experimental results show that our proposed model obtains obvious performance gain over several strong baselines, and further human evaluation validates that our model can ask questions of high quality by leveraging paraphrase knowledge.

NeuInfer: Knowledge Inference on N-ary Facts [Website][PDF]
Saiping Guan, Xiaolong Jin, Jiafeng Guo, Yuanzhuo Wang, and Xueqi Cheng 5:00–6:00

Knowledge inference on knowledge graph has attracted extensive attention, which aims to find out connotative valid facts in knowledge graph and is very helpful for improving the performance of many downstream applications. However, researchers have mainly poured attention to knowledge inference on binary facts. The studies on n-ary facts are relatively scarcer, although they are also ubiquitous in the real world. Therefore, this paper addresses knowledge inference on n-ary facts. We represent each n-ary fact as a primary triple coupled with a set of its auxiliary descriptive attribute-value pair(s). We further propose a neural network model, NeuInfer, for knowledge inference on n-ary facts. Besides handling the common task to infer an unknown element in a whole fact, NeuInfer can cope with a new type of task, flexible knowledge inference. It aims to infer an unknown element in a partial fact consisting of the primary triple coupled with any number of its auxiliary description(s). Experimental results demonstrate the remarkable superiority of NeuInfer.

Neural Graph Matching Networks for Chinese Short Text Matching [Website][PDF]
Lu Chen, Yanbin Zhao, Boer Lyu, Lesheng Jin, Zhi Chen, Su Zhu, and Kai Yu 5:00–6:00

Chinese short text matching usually employs word sequences rather than character sequences to get better performance. However, Chinese word segmentation can be erroneous, ambiguous or inconsistent, which consequently hurts the final matching performance. To address this problem, we propose neural graph matching networks, a novel sentence matching framework capable of dealing with multi-granular input information. Instead of a character sequence or a single word sequence, paired word lattices formed from multiple word segmentation hypotheses are used as input and the model learns a graph representation according to an attentive graph matching mechanism. Experiments on two Chinese datasets show that our models outperform the state-of-the-art short text matching models.

Neural Mixed Counting Models for Dispersed Topic Discovery [Website][PDF]
Jiemin Wu, Yanghui Rao, Zusheng Zhang, Haoran Xie, Qing Li, Fu Lee Wang, and Ziye Chen 5:00–6:00

Mixed counting models that use the negative binomial distribution as the prior can well model over-dispersed and hierarchically dependent random variables; thus they have attracted much attention in mining dispersed document topics. However, the existing parameter inference method like Monte Carlo sampling is quite time-consuming. In this paper, we propose two efficient neural mixed counting models, i.e., the Negative Binomial-Neural Topic Model (NB-NTM) and the Gamma Negative Binomial-Neural Topic Model (GNB-NTM) for dispersed topic discovery. Neural variational inference algorithms are developed to infer model parameters by using the reparameterization of Gamma distribution and the Gaussian approximation of Poisson distribution. Experiments on real-world datasets indicate that our models outperform state-of-the-art baseline models in terms of perplexity and topic coherence. The results also validate that both NB-NTM and GNB-NTM can produce explainable intermediate variables by generating dispersed proportions of document topics.

Reasoning Over Semantic-Level Graph for Fact Checking [Website][PDF]
Wanjun Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin
5:00–6:00

Fact checking is a challenging task because verifying the truthfulness of a claim requires reasoning about multiple retrievable evidence. In this work, we present a method suitable for reasoning about the semantic-level structure of evidence. Unlike most previous works, which typically represent evidence sentences with either string concatenation or fusing the features of isolated evidence sentences, our approach operates on rich semantic structures of evidence obtained by semantic role labeling. We propose two mechanisms to exploit the structure of evidence while leveraging the advances of pre-trained models like BERT, GPT or XLNet. Specifically, using XLNet as the backbone, we first utilize the graph structure to re-define the relative distances of words, with the intuition that semantically related words should have short distances. Then, we adopt graph convolutional network and graph attention network to propagate and aggregate information from neighboring nodes on the graph. We evaluate our system on FEVER, a benchmark dataset for fact checking, and find that rich structural information is helpful and both our graph-based mechanisms improve the accuracy. Our model is the state-of-the-art system in terms of both official evaluation metrics, namely claim verification accuracy and FEVER score.

Session 11A: Student Research Workshop

HGCN4MeSH: Hybrid Graph Convolution Network for MeSH Indexing [Website][PDF]
Miaomiao Yu, Yujiu Yang, and Chenhui Li 5:00–6:00

Recently deep learning has been used in Medical subject headings (MeSH) indexing to reduce the time and monetary cost by manual annotation, including DeepMeSH, TextCNN, etc. However, these models still suffer from failing to capture the complex correlations between MeSH terms. To this end, we introduce Graph Convolution Network (GCN) to learn the relationship between these terms, and present a novel Hybrid Graph Convolution Net for MeSH index (HGCN4MeSH). Basically, we utilize two BiGRUs to learn the embedding representation of the abstract and the title of the MeSH index text respectively. At the same time, we establish the adjacency matrix of MeSH terms based on the co-occurrence relationships in Corpus, which is easy to apply for GCN representation learning. On the basis of learning the mixed representation, the prediction problem of the MeSH index keywords is transformed into an extreme multi-label classification problem after the attention layer operation. Experimental results on two datasets show that HGCN4MeSH is competitive compared with the state-of-the-art methods.

Considering Likelihood in NLP Classification Explanations with Occlusion and Language Modeling [Website][PDF]
David Harbecke and Christoph Alt 5:00–6:00

Recently, state-of-the-art NLP models gained an increasing syntactic and semantic understanding of language, and explanation methods are crucial to understand their decisions. Occlusion is a well established method that provides explanations on discrete language data, e.g. by removing a language unit from an input and measuring the impact on a model's decision. We argue that current occlusion-based methods often produce invalid or syntactically incorrect language data, neglecting the improved abilities of recent NLP models. Furthermore, gradient-based explanation methods disregard the discrete distribution of data in NLP. Thus, we propose OLM: a novel explanation method that combines occlusion and language models to sample valid and syntactically correct replacements with high likelihood, given the context of the original input. We lay out a theoretical foundation that alleviates these weaknesses of other explanation methods in NLP and provide results that underline the importance of considering data likelihood in occlusion-based explanation.

Session 11A: Summarization-6

Automatic Generation of Citation Texts in Scholarly Papers: A Pilot Study [Website][PDF]
Xinyu Xing, Xiaosheng Fan, and Xiaojun Wan 5:00–6:00

In this paper, we study the challenging problem of automatic generation of citation texts in scholarly papers. Given the context of a citing paper A and a cited paper B, the task aims to generate a short text to describe B in the given context of A. One big challenge for addressing this task is the lack of training data. Usually, explicit citation texts are easy to extract, but it is not easy to extract implicit citation texts from scholarly papers. We thus first train an implicit citation extraction model based on BERT and leverage the model to construct a large training dataset for the citation text generation task. Then we propose and train a multi-source pointer-generator network with cross attention mechanism for citation text generation. Empirical evaluation results on a manually labeled test dataset verify the efficacy of our model. This pilot study confirms the feasibility of automatically generating citation texts in scholarly papers and the technique has the great potential to help researchers prepare their scientific papers.

Composing Elementary Discourse Units in Abstractive Summarization [Website][PDF]
Zhenwen Li, Wenhao Wu, and Sujian Li 5:00–6:00

In this paper, we argue that elementary discourse unit (EDU) is a more appropriate textual unit of content selection than the sentence unit in abstractive summarization. To well handle the problem of composing EDUs into an informative and fluent summary, we propose a novel summarization method that first designs an EDU selection model to extract and group informative EDUs and then an EDU fusion model to fuse the EDUs in each group into one sentence. We also design the reinforcement learning mechanism to use EDU fusion results to reward the EDU selection action, boosting the final summarization performance. Experiments on CNN/Daily Mail have demonstrated the effectiveness of our model.

Extractive Summarization as Text Matching [Website][PDF]
Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang 5:00–6:00

This paper creates a paradigm shift with regard to the way we build neural extractive summarization systems. Instead of following the commonly used framework of extracting sentences individually and modeling the relationship between sentences, we formulate the extractive summarization task as a semantic text matching problem, in which a source document and candidate summaries will be (extracted from the original text) matched in a semantic space. Notably, this paradigm shift to semantic matching framework is well-grounded in our comprehensive analysis of the inherent gap between sentence-level and summary-level extractors based on the property of the dataset. Besides, even instantiating the framework with a simple form of a matching model, we have driven the state-of-the-art extractive result on CNN/DailyMail to a new level (44.41 in ROUGE-1). Experiments on the other five datasets also show the effectiveness of the matching framework. We believe the power of this matching-based summarization framework has not been fully exploited. To encourage more instantiations in the future, we have released our codes, processed dataset, as well as generated summaries in `https://github.com/maszhongming/MatchSum`.

Heterogeneous Graph Neural Networks for Extractive Document Summarization [Website][PDF]
Danqing Wang, Pengfei Liu, Yining Zheng, Xipeng Qiu, and Xuanjing Huang 5:00–6:00

As a crucial step in extractive document summarization, learning cross-sentence relations has been explored by a plethora of approaches. An intuitive way is to put them in the graph-based neural network, which has a more complex structure for capturing inter-sentence relationships. In this paper, we present a heterogeneous graph-based neural network for extractive summarization (HETERSUMGRAPH), which contains semantic nodes of different granularity levels apart from sentences. These additional nodes act as the intermediary between sentences and enrich the cross-sentence relations. Besides, our graph structure is flexible in natural extension from a single-document setting to multi-document via introducing document nodes. To our knowledge, we are the first one to introduce different types of nodes into graph-based neural networks for extractive document summarization and perform a comprehensive qualitative analysis to investigate their benefits. The code will be released on Github.

Jointly Learning to Align and Summarize for Neural Cross-Lingual Summarization [Website][PDF]
Yue Cao, Hui Liu, and Xiaojun Wan 5:00–6:00

Cross-lingual summarization is the task of generating a summary in one language given a text in a different language. Previous works on cross-lingual summarization mainly focus on using pipeline methods or training an end-to-end model using the translated parallel data. However, it is a big challenge for the model to directly learn cross-lingual summarization as it requires learning to understand different languages and learning how to summarize at the same time. In this paper, we propose to ease the cross-lingual summarization training by jointly learning to align and summarize. We design relevant loss functions to train this framework and propose several methods to enhance the isomorphism and cross-lingual transfer between languages. Experimental results show that our model can outperform competitive models in most cases. In addition, we show that our model even has the ability to generate cross-lingual summaries without access to any cross-lingual corpus.

Leveraging Graph to Improve Abstractive Multi-Document Summarization [Website][PDF]
Wei Li, Xinyan Xiao, Jiachen Liu, Hua Wu, Haifeng Wang, and Junping Du 5:00–6:00

Graphs that capture relations between textual units have great benefits for detecting salient information from multiple documents and generating overall coherent summaries. In this paper, we develop a neural abstractive multi-document summarization (MDS) model which can leverage well-known graph representations of documents such as similarity graph and discourse graph, to more effectively process multiple input documents and produce abstractive summaries. Our model utilizes graphs to encode documents in order to capture cross-document relations, which is

crucial to summarizing long documents. Our model can also take advantage of graphs to guide the summary generation process, which is beneficial for generating coherent and concise summaries. Furthermore, pre-trained language models can be easily combined with our model, which further improve the summarization performance significantly. Empirical results on the WikiSum and MultiNews dataset show that the proposed architecture brings substantial improvements over several strong baselines.

Multi-Granularity Interaction Network for Extractive and Abstractive Multi-Document Summarization [Website] [PDF]

Hanqi Jin, Tianming Wang, and Xiaojun Wan 5:00–6:00

In this paper, we propose a multi-granularity interaction network for extractive and abstractive multi-document summarization, which jointly learn semantic representations for words, sentences, and documents. The word representations are used to generate an abstractive summary while the sentence representations are used to produce an extractive summary. We employ attention mechanisms to interact between different granularity of semantic representations, which helps to capture multi-granularity key information and improves the performance of both abstractive and extractive summarization. Experiment results show that our proposed model substantially outperforms all strong baseline methods and achieves the best results on the Multi-News dataset.

Session 11A Syntax: Tagging, Chunking and Parsing-4

Tetra-Tagging: Word-Synchronous Parsing with Linear-Time Inference [Website][PDF]
Nikita Kitaev and Dan Klein 5:00–6:00

We present a constituency parsing algorithm that, like a supertagger, works by assigning labels to each word in a sentence. In order to maximally leverage current neural architectures, the model scores each word's tags in parallel, with minimal task-specific structure. After scoring, a left-to-right reconciliation phase extracts a tree in (empirically) linear time. Our parser achieves 95.4 F1 on the WSJ test set while also achieving substantial speedups compared to current state-of-the-art parsers with comparable accuracies.

Session 11A: Theme-3

Are we Estimating or Guesstimating Translation Quality? [Website][PDF]
Shuo Sun, Francisco Guzmán, and Lucia Specia 5:00–6:00

Recent advances in pre-trained multilingual language models lead to state-of-the-art results on the task of quality estimation (QE) for machine translation. A carefully engineered ensemble of such models won the QE shared task at WMT19. Our in-depth analysis, however, shows that the success of using pre-trained language models for QE is over-estimated due to three issues we observed in current QE datasets: (i) The distributions of quality scores are imbalanced and skewed towards good quality scores; (iii) QE models can perform well on these datasets while looking at only source or translated sentences; (iii) They contain statistical artifacts that correlate well with human-annotated QE labels. Our findings suggest that although QE models might capture fluency of translated sentences and complexity of source sentences, they cannot model adequacy of translations effectively.

Language (Re)modelling: Towards Embodied Language Understanding [Website][PDF]
Ronen Tamari, Chen Shani, Tom Hope, Miriam R L Petruck, Omri Abend, and Dafna Shahaf 5:00–6:00

While natural language understanding (NLU) is advancing rapidly, today's technology differs from human-like language understanding in fundamental ways, notably in its inferior efficiency, interpretability, and generalization. This work proposes an approach to representation and learning based on the tenets of embodied cognitive linguistics (ECL). According to ECL, natural language is inherently executable (like programming languages), driven by mental simulation and metaphoric mappings over hierarchical compositions of structures and schemata learned through embodied interaction. This position paper argues that the use of grounding by metaphoric reasoning and simulation will greatly benefit NLU systems, and proposes a system architecture along with a roadmap towards realizing this vision.

The State and Fate of Linguistic Diversity and Inclusion in the NLP World [Website][PDF]
Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury 5:00–6:00

Language technologies contribute to promoting multilingualism and linguistic diversity around the world. However, only a very small number of the over 7000 languages of the world are represented in the rapidly evolving language technologies and applications. In this paper we look at the relation between the types of languages, resources, and their representation in NLP conferences to understand the trajectory that different languages have followed over time. Our quantitative investigation underlines the disparity between languages, especially in terms of their resources, and calls into question the "language agnostic" status of current models and systems. Through this paper, we attempt to convince the ACL community to prioritise the resolution of the predicaments highlighted here, so that no language is left behind.

The Unstoppable Rise of Computational Linguistics in Deep Learning [Website][PDF]
James Henderson 5:00–6:00

In this paper, we trace the history of neural networks applied to natural language understanding tasks, and identify key contributions which the nature of language has made to the development of neural network architectures. We focus on the importance of variable binding and its instantiation in attention-based models, and argue that Transformer is not a sequence model but an induced-structure model. This perspective leads to predictions of the challenges facing research in deep learning architectures for natural language understanding.

To Boldly Query What No One Has Annotated Before? The Frontiers of Corpus Querying [Website][PDF]
Markus Gärtner and Kerstin Jung 5:00–6:00

Corpus query systems exist to address the multifarious information needs of any person interested in the content of annotated corpora. In this role they play an important part in making those resources usable for a wider audience. Over the past decades, several such query systems and languages have emerged, varying greatly in their expressiveness and technical details. This paper offers a broad overview of the history of corpora and corpus query tools. It focusses strongly on the query side and hints at exciting directions for future development.

Demo Session 1B

Time: *5:45–6:30*

Clinical-Coder: Assigning Interpretable ICD-10 Codes to Chinese Clinical Notes [Website][PDF]
Pengfei Cao, Chenwei Yan, Xiangling Fu, Yubo Chen, Kang Liu, Jun Zhao, Shengping Liu, and Weifeng Chong

In this paper, we introduce Clinical-Coder, an online system aiming to assign ICD codes to Chinese clinical notes. ICD coding has been a research hotspot of clinical medicine, but the interpretability of prediction hinders its practical application. We exploit a Dilated Convolutional Attention network with N-gram Matching mechanism (DCANM) to capture semantic features for non-continuous words and continuous n-gram words, concentrating on explaining the reason why each ICD code to be predicted. The experiments demonstrate that our approach is effective and that our system is able to provide supporting information in clinical decision making.

ESPnet-ST: All-in-One Speech Translation Toolkit [Website][PDF]
Hirofumi Inaguma, Shun Kiyono, Kevin Duh, Shigeki Karita, Nelson Yalta, Tomoki Hayashi, and Shinji Watanabe

We present ESPnet-ST, which is designed for the quick development of speech-to-speech translation systems in a single framework. ESPnet-ST is a new project inside end-to-end speech processing toolkit, ESPnet, which integrates or newly implements automatic speech recognition, machine translation, and text-to-speech functions for speech translation. We provide all-in-one recipes including data pre-processing, feature extraction, training, and decoding pipelines for a wide range of benchmark datasets. Our reproducible results can match or even outperform the current state-of-the-art performances; these pre-trained models are downloadable. The toolkit is publicly available at https://github.com/espnet/espnet.

Session 11B Overview – Wednesday, July 8, 2020 6:00–7:00

Track A *Dialogue and Interactive Systems-14* Abstracts	A Contextual Hierarchical Attention Network with Adaptive Objective for Dialogue State Tracking *Shan, Li, Zhang, Meng, Feng, Niu, and Zhou* [Website] [PDF]	Data Manipulation: Towards Effective Instance Learning for Neural Dialogue Generation via Learning to Augment and Reweight *Cai, Chen, Song, Zhang, Zhao, and Yin* [Website] [PDF]	Dynamic Fusion Network for Multi-Domain End-to-end Task-Oriented Dialog *Qin, Xu, Che, Zhang, and Liu* [Website] [PDF]	Learning Efficient Dialogue Policy from Demonstrations through Shaping *Wang, Peng, and Wong* [Website] [PDF]	SAS: Dialogue State Tracking via Slot Attention and Slot Information Sharing *Hu, Yang, Chen, and Yu* [Website] [PDF]
	Speaker Sensitive Response Evaluation Model *Bak and Oh* [Website] [PDF]				
Track B *Discourse and Pragmatics-6* Abstracts	A Top-down Neural Architecture towards Text-level Parsing of Discourse Rhetorical Structure *Zhang, Xing, Kong, Li, and Zhou* [Website] [PDF]	[TACL] Unsupervised Discourse Constituency Parsing Using Viterbi EM *Nishida and Nakayama* [Website] [PDF]			
Track C *Information Extraction-4* Abstracts	Amalgamation of protein sequence, structure and textual information for improving protein-protein interaction identification *Dutta and Saha* [Website] [PDF]	Bipartite Flat-Graph Network for Nested Named Entity Recognition *Luo and Zhao* [Website] [PDF]	Connecting Embeddings for Knowledge Graph Entity Typing *Zhao, Xie, Liu, and WANG* [Website] [PDF]	Continual Relation Learning via Episodic Memory Activation and Reconsolidation *Han, Dai, Gao, Lin, Liu, Li, Sun, and Zhou* [Website] [PDF]	Handling Rare Entities for Neural Sequence Labeling *Li, Li, Yao, and Li* [Website] [PDF]
	Instance-Based Learning of Span Representations: A Case Study through Named Entity Recognition *Ouchi, Suzuki, Kobayashi, Yokoi, Kuribayashi, Konno, and Inui* [Website] [PDF]	MIE: A Medical Information Extractor towards Medical Dialogues *Zhang, Jiang, Zhang, Liu, Cao, Liu, Liu, and Zhao* [Website] [PDF]	Named Entity Recognition as Dependency Parsing *Yu, Bohnet, and Poesio* [Website] [PDF]	Neighborhood Matching Network for Entity Alignment *Wu, Liu, Feng, Wang, and Zhao* [Website] [PDF]	Relation Extraction with Explanation *Shahbazi, Fern, Ghaeini, and Tadepalli* [Website] [PDF]
	Representation Learning for Information Extraction from Form-like Documents *Majumder, Potti, Tata, Wendt, Zhao, and Najork* [Website] [PDF]	Single-/Multi-Source Cross-Lingual NER via Teacher-Student Learning on Unlabeled Data in Target Language *Wu, Lin, Karlsson, LOU, and Huang* [Website] [PDF]	Synchronous Double-channel Recurrent Network for Aspect-Opinion Pair Extraction *Chen, Liu, Wang, Zhang, and Chi* [Website] [PDF]		

Track D *Language Grounding to Vision, Robotics and Beyond-5* Abstracts	Cross-modal Coherence Modeling for Caption Generation *Alikhani, Sharma, Li, Soricut, and Stone* [Website][PDF]	Knowledge Supports Visual Language Grounding: A Case Study on Colour Terms *Schüz and Zarrieß* [Website][PDF]	Span-based Localizing Network for Natural Language Video Localization *Zhang, Sun, Jing, and Zhou* [Website][PDF]	Words Aren't Enough, Their Order Matters: On the Robustness of Grounding Visual Referring Expressions *Akula, Gella, Al-Onaizan, Zhu, and Reddy* [Website][PDF]	
Track E *Machine Learning for NLP-12* Abstracts	A Mixture of h - 1 Heads is Better than h Heads *Peng, Schwartz, Li, and Smith* [Website][PDF]	Dependency Graph Enhanced Dual-transformer Structure for Aspect-based Sentiment Classification *Tang, Ji, Li, and Zhou* [Website][PDF]	Differentiable Window for Dynamic Local Attention *Nguyen, Nguyen, Joty, and Li* [Website][PDF]	Evaluating and Enhancing the Robustness of Neural Network-based Dependency Parsing Models with Adversarial Examples *Zheng, Zeng, Zhou, Hsieh, Cheng, and Huang* [Website][PDF]	Exploiting Syntactic Structure for Better Language Modeling: A Syntactic Distance Approach *Du, Lin, Shen, O'Donnell, Bengio, and Zhang* [Website][PDF]
	Learning Architectures from an Extended Search Space for Language Modeling *Li, Hu, Zhang, Xu, Jiang, Xiao, Zhu, Liu, and* [Website][PDF]	The Right Tool for the Job: Matching Model and Instance Complexities *Schwartz, Stanovsky, Swayamdipta, Dodge, and Smith* [Website][PDF]			
Track F *Phonology, Morphology and Word Segmentation-3* Abstracts	Bootstrapping Techniques for Polysynthetic Morphological Analysis *Lane and Bird* [Website][PDF]	Coupling Distant Annotation and Adversarial Training for Cross-Domain Chinese Word Segmentation *Ding, Long, Xu, Zhu, Xie, Wang, and Zheng* [Website][PDF]	Modeling Morphological Typology for Unsupervised Learning of Language Morphology *Xu, Kodner, Marcus, and Yang* [Website][PDF]	Predicting Declension Class from Form and Meaning *Williams, Pimentel, Blix, McCarthy, Chodroff, and Cotterell* [Website][PDF]	Unsupervised Morphological Paradigm Completion *Jin, Cai, Peng, Xia, McCarthy, and Kann* [Website][PDF]
Track G *Question Answering-9* Abstracts	Document Modeling with Graph Attention Networks for Multi-grained Machine Reading Comprehension *Zheng, Wen, Liang, Duan, Che, Jiang, Zhou, and Liu* [Website][PDF]	Harvesting and Refining Question-Answer Pairs for Unsupervised QA *Li, Wang, Dong, Wei, and Xu* [Website][PDF]	Low-Resource Generation of Multi-hop Reasoning Questions *Yu, Liu, Qiu, Su, Wang, Quan, and Yin* [Website][PDF]	R4C: A Benchmark for Evaluating RC Systems to Get the Right Answer for the Right Reason *Inoue, Stenetorp, and Inui* [Website][PDF]	Recurrent Chunking Mechanisms for Long-Text Machine Reading Comprehension *Gong, Shen, Yu, Chen, and Yu* [Website][PDF]
	RikiNet: Reading Wikipedia Pages for Natural Question Answering *Liu, Gong, Fu, Yan, Chen, Jiang, Lv, and Duan* [Website][PDF]				

Track H *Sentence Level-6* Abstracts	[TACL] AMR-To-Text Generation with Graph Transformer *Wang, Wan, and Jin* [Website][PDF]	Parsing into Variable-in-situ Logico-Semantic Graphs *Chen and Sun* [Website][PDF]	Semantic Parsing for English as a Second Language *Zhao, Sun, and Wan* [Website][PDF]	Semi-Supervised Semantic Dependency Parsing Using CRF Autoencoders *Jia, Ma, Cai, and Tu* [Website][PDF]	Unsupervised Dual Paraphrasing for Two-stage Semantic Parsing *Cao, Zhu, Yang, Liu, Ma, Zhao, Chen, and Yu* [Website][PDF]
Track I *Student Research Workshop* Abstracts	Feature Difference Makes Sense: A medical image captioning model exploiting feature difference and tag information *Park, Kim, Yoon, Park, and Choi* [Website][PDF]	Multi-Task Neural Model for Agglutinative Language Translation *Pan, Li, Yang, and Dong* [Website][PDF]			

Session 11B Details

Session 11B: Dialogue and Interactive Systems-14

A Contextual Hierarchical Attention Network with Adaptive Objective for Dialogue State Tracking [Website][PDF]
Yong Shan, Zekang Li, Jinchao Zhang, Fandong Meng, Yang Feng, Cheng Niu, and Jie Zhou 6:00–7:00

Recent studies in dialogue state tracking (DST) leverage historical information to determine states which are generally represented as slot-value pairs. However, most of them have limitations to efficiently exploit relevant context due to the lack of a powerful mechanism for modeling interactions between the slot and the dialogue history. Besides, existing methods usually ignore the slot imbalance problem and treat all slots indiscriminately, which limits the learning of hard slots and eventually hurts overall performance. In this paper, we propose to enhance the DST through employing a contextual hierarchical attention network to not only discern relevant information at both word level and turn level but also learn contextual representations. We further propose an adaptive objective to alleviate the slot imbalance problem by dynamically adjust weights of different slots during training. Experimental results show that our approach reaches 52.68% and 58.55% joint accuracy on MultiWOZ 2.0 and MultiWOZ 2.1 datasets respectively and achieves new state-of-the-art performance with considerable improvements (+1.24% and +5.98%).

Data Manipulation: Towards Effective Instance Learning for Neural Dialogue Generation via Learning to Augment and Reweight [Website][PDF]
Hengyi Cai, Hongshen Chen, Yonghao Song, Cheng Zhang, Xiaofang Zhao, and Dawei Yin 6:00–7:00

Current state-of-the-art neural dialogue models learn from human conversations following the data-driven paradigm. As such, a reliable training corpus is the crux of building a robust and well-behaved dialogue model. However, due to the open-ended nature of human conversations, the quality of user-generated training data varies greatly, and effective training samples are typically insufficient while noisy samples frequently appear. This impedes the learning of those data-driven neural dialogue models. Therefore, effective dialogue learning requires not only more reliable learning samples, but also fewer noisy samples. In this paper, we propose a data manipulation framework to proactively reshape the data distribution towards reliable samples by augmenting and highlighting effective learning samples as well as reducing the effect of inefficient samples simultaneously. In particular, the data manipulation model selectively augments the training samples and assigns an importance weight to each instance to reform the training data. Note that, the proposed data manipulation framework is fully data-driven and learnable. It not only manipulates training samples to optimize the dialogue generation model, but also learns to increase its manipulation skills through gradient descent with validation samples. Extensive experiments show that our framework can improve the dialogue generation performance with respect to various automatic evaluation metrics and human judgments.

Dynamic Fusion Network for Multi-Domain End-to-end Task-Oriented Dialog [Website][PDF]
Libo Qin, Xiao Xu, Wanxiang Che, Yue Zhang, and Ting Liu 6:00–7:00

Recent studies have shown remarkable success in end-to-end task-oriented dialog system. However, most neural models rely on large training data, which are only available for a certain number of task domains, such as navigation and scheduling. This makes it difficult to scalable for a new domain with limited labeled data. However, there has been relatively little research on how to effectively use data from all domains to improve the performance of each domain and also unseen domains. To this end, we investigate methods that can make explicit use of domain knowledge and introduce a shared-private network to learn shared and specific knowledge. In addition, we propose a novel Dynamic Fusion Network (DF-Net) which automatically exploit the relevance between the target domain and each domain. Results show that our models outperforms existing methods on multi-domain dialogue, giving the state-of-the-art in the literature. Besides, with little training data, we show its transferability by outperforming prior best model by 13.9% on average.

Learning Efficient Dialogue Policy from Demonstrations through Shaping [Website][PDF]
Huimin Wang, Baolin Peng, and Kam-Fai Wong 6:00–7:00

Training a task-oriented dialogue agent with reinforcement learning is prohibitively expensive since it requires a large volume of interactions with users. Human demonstrations can be used to accelerate learning progress. However, how to effectively leverage demonstrations to learn dialogue policy remains less explored. In this paper, we present S^2Agent that efficiently learns dialogue policy from demonstrations through policy shaping and reward shaping. We use an imitation model to distill knowledge from demonstrations, based on which policy shaping estimates feedback on how the agent should act in policy space. Reward shaping is then incorporated to bonus state-actions similar to demonstrations explicitly in value space encouraging better exploration. The effectiveness of the proposed S^2Agentt is demonstrated in three dialogue domains and a challenging domain adaptation task with both user simulator evaluation and human evaluation.

SAS: Dialogue State Tracking via Slot Attention and Slot Information Sharing [Website][PDF]
Jiaying Hu, Yan Yang, Chencai Chen, liang he liang, and Zhou Yu 6:00–7:00

Dialogue state tracker is responsible for inferring user intentions through dialogue history. Previous methods have difficulties in handling dialogues with long interaction context, due to the excessive information. We propose a Dialogue State Tracker with Slot Attention and Slot Information Sharing (SAS) to reduce redundant information's interference and improve long dialogue context tracking. Specially, we first apply a Slot Attention to learn a set of slot-specific fea-

tures from the original dialogue and then integrate them using a slot information sharing module. Our model yields a significantly improved performance compared to previous state-of the-art models on the MultiWOZ dataset.

Speaker Sensitive Response Evaluation Model [Website][PDF]
JinYeong Bak and Alice Oh 6:00–7:00

Automatic evaluation of open-domain dialogue response generation is very challenging because there are many appropriate responses for a given context. Existing evaluation models merely compare the generated response with the ground truth response and rate many of the appropriate responses as inappropriate if they deviate from the ground truth. One approach to resolve this problem is to consider the similarity of the generated response with the conversational context. In this paper, we propose an automatic evaluation model based on that idea and learn the model parameters from an unlabeled conversation corpus. Our approach considers the speakers in defining the different levels of similar context. We use a Twitter conversation corpus that contains many speakers and conversations to test our evaluation model. Experiments show that our model outperforms the other existing evaluation metrics in terms of high correlation with human annotation scores. We also show that our model trained on Twitter can be applied to movie dialogues without any additional training. We provide our code and the learned parameters so that they can be used for automatic evaluation of dialogue response generation models.

Session 11B: Discourse and Pragmatics-6

A Top-down Neural Architecture towards Text-level Parsing of Discourse Rhetorical Structure [Website][PDF]
Longyin Zhang, Yuqing Xing, Fang Kong, Peifeng Li, and Guodong Zhou 6:00–7:00

Due to its great importance in deep natural language understanding and various down-stream applications, text-level parsing of discourse rhetorical structure (DRS) has been drawing more and more attention in recent years. However, all the previous studies on text-level discourse parsing adopt bottom-up approaches, which much limit the DRS determination on local information and fail to well benefit from global information of the overall discourse. In this paper, we justify from both computational and perceptive points-of-view that the top-down architecture is more suitable for text-level DRS parsing. On the basis, we propose a top-down neural architecture toward text-level DRS parsing. In particular, we cast discourse parsing as a recursive split point ranking task, where a split point is classified to different levels according to its rank and the elementary discourse units (EDUs) associated with it are arranged accordingly. In this way, we can determine the complete DRS as a hierarchical tree structure via an encoder-decoder with an internal stack. Experimentation on both the English RST-DT corpus and the Chinese CDTB corpus shows the great effectiveness of our proposed top-down approach towards text-level DRS parsing.

[TACL] Unsupervised Discourse Constituency Parsing Using Viterbi EM [Website][PDF]
Noriki Nishida and Hideki Nakayama 6:00–7:00

In this paper, we introduce an unsupervised discourse constituency parsing algorithm. We use Viterbi EM with a margin-based criterion to train a span-based discourse parser in an unsupervised manner. We also propose initialization methods for Viterbi training of discourse constituents based on our prior knowledge of text structures. Experimental results demonstrate that our unsupervised parser achieves comparable or even superior performance to fully supervised parsers. We also investigate discourse constituents that are learned by our method.

Session 11B: Information Extraction-4

Amalgamation of protein sequence, structure and textual information for improving protein-protein interaction identification [Website][PDF]
pratik Dutta and Sriparna Saha 6:00–7:00

An in-depth exploration of protein-protein interactions (PPI) is essential to understand the metabolism in addition to the regulations of biological entities like proteins, carbohydrates, and many more. Most of the recent PPI tasks in BioNLP domain have been carried out solely using textual data. In this paper, we argue that incorporating multimodal cues can improve the automatic identification of PPI. As a first step towards enabling the development of multimodal approaches for PPI identification, we have developed two multi-modal datasets which are extensions and multi-modal versions of two popular benchmark PPI corpora (BioInfer and HRPD50). Besides, existing textual modalities, two new modalities, 3D protein structure and underlying genomic sequence, are also added to each instance. Further, a novel deep multi-modal architecture is also implemented to efficiently predict the protein interactions from the developed datasets. A detailed experimental analysis reveals the superiority of the multi-modal approach in comparison to the strong baselines including unimodal approaches and state-of-the-art methods over both the generated multi-modal datasets. The developed multi-modal datasets are available for use at https://github.com/sduttap16/MM_PPI_NLP.

Bipartite Flat-Graph Network for Nested Named Entity Recognition [Website][PDF]
Ying Luo and Hai Zhao 6:00–7:00

In this paper, we propose a novel bipartite flat-graph network (BiFlaG) for nested named entity recognition (NER), which contains two subgraph modules: a flat NER module for outermost entities and a graph module for all the entities located in inner layers. Bidirectional LSTM (BiLSTM) and graph convolutional network (GCN) are adopted to jointly learn flat entities and their inner dependencies. Different from previous models, which only consider the unidirectional delivery of information from innermost layers to outer ones (or outside-to-inside), our model effectively captures the bidirectional interaction between them. We first use the entities recognized by the flat NER module to construct an entity graph, which is fed to the next graph module. The richer representation learned from graph module carries the dependencies of inner entities and can be exploited to improve outermost entity predictions. Experimental results on three standard nested NER datasets demonstrate that our BiFlaG outperforms previous state-of-the-art models.

Connecting Embeddings for Knowledge Graph Entity Typing [Website][PDF]
Yu Zhao, anxiang zhang anxiang, Ruobing Xie, Kang Liu, and Xiaojie WANG 6:00–7:00

Knowledge graph (KG) entity typing aims at inferring possible missing entity type instances in KG, which is a very significant but still under-explored subtask of knowledge graph completion. In this paper, we propose a novel approach for KG entity typing which is trained by jointly utilizing local typing knowledge from existing entity type assertions and global triple knowledge in KGs. Specifically, we present two distinct knowledge-driven effective mechanisms of entity type inference. Accordingly, we build two novel embedding models to realize the mechanisms. Afterward, a joint model via connecting them is used to infer missing entity type instances, which favors inferences that agree with both entity type instances and triple knowledge in KGs. Experimental results on two real-world datasets (Freebase and YAGO) demonstrate the effectiveness of our proposed mechanisms and models for improving KG entity typing. The source code and data of this paper can be obtained from: https://github.com/Adam1679/ConnectE .

Continual Relation Learning via Episodic Memory Activation and Reconsolidation [Website][PDF]
Xu Han, Yi Dai, Tianyu Gao, Yankai Lin, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou 6:00–7:00

Continual relation learning aims to continually train a model on new data to learn incessantly emerging novel relations while avoiding catastrophically forgetting old relations. Some pioneering work has proved that storing a handful of historical relation examples in episodic memory and replaying them in subsequent training is an effective solution for such a challenging problem. However, these memory-based methods usually suffer from overfitting the few memorized examples of old relations, which may gradually cause inevitable confusion among existing relations. Inspired by the mechanism in human long-term memory formation, we introduce episodic memory activation and reconsolidation (EMAR) to continual relation learning. Every time neural models are activated to learn both new and memorized data, EMAR utilizes relation prototypes for memory reconsolidation exercise to keep a stable understanding of old relations. The experimental results show that EMAR could get rid of catastrophically forgetting old relations and outperform the state-of-the-art continual learning models.

Handling Rare Entities for Neural Sequence Labeling [Website][PDF]
Yangming Li, Han Li, Kaisheng Yao, and Xiaolong Li 6:00–7:00

One great challenge in neural sequence labeling is the data sparsity problem for rare entity words and phrases. Most of test set entities appear only few times and are even unseen in training corpus, yielding large number of out-of-vocabulary (OOV) and low-frequency (LF) entities during evaluation. In this work, we propose approaches to address this problem. For OOV entities, we introduce local context reconstruction to implicitly incorporate contextual information into their representations. For LF entities, we present delexicalized entity identification to explicitly extract their frequency-agnostic and entity-type-specific representations. Extensive experiments on multiple benchmark datasets show that our model has significantly outperformed all previous methods and achieved new start-of-the-art results. Notably, our methods surpass the model fine-tuned on pre-trained language models without external resource.

Instance-Based Learning of Span Representations: A Case Study through Named Entity Recognition
[Website][PDF]
Hiroki Ouchi, Jun Suzuki, Sosuke Kobayashi, Sho Yokoi, Tatsuki Kuribayashi, Ryuto Konno, and Kentaro Inui 6:00–7:00

Interpretable rationales for model predictions play a critical role in practical applications. In this study, we develop models possessing interpretable inference process for structured prediction. Specifically, we present a method of instance-based learning that learns similarities between spans. At inference time, each span is assigned a class label based on its similar spans in the training set, where it is easy to understand how much each training instance contributes to the predictions. Through empirical analysis on named entity recognition, we demonstrate that our method enables to build models that have high interpretability without sacrificing performance.

MIE: A Medical Information Extractor towards Medical Dialogues [Website][PDF]
Yuanzhe Zhang, Zhongtao Jiang, Tao Zhang, Shiwan Liu, Jiarun Cao, Kang Liu, Shengping Liu, and Jun Zhao 6:00–7:00

Electronic Medical Records (EMRs) have become key components of modern medical care systems. Despite the merits of EMRs, many doctors suffer from writing them, which is time-consuming and tedious. We believe that automatically converting medical dialogues to EMRs can greatly reduce the burdens of doctors, and extracting information from medical dialogues is an essential step. To this end, we annotate online medical consultation dialogues in a window-sliding style, which is much easier than the sequential labeling annotation. We then propose a Medical Information Extractor (MIE) towards medical dialogues. MIE is able to extract mentioned symptoms, surgeries, tests, other information and their corresponding status. To tackle the particular challenges of the task, MIE uses a deep matching architecture, taking dialogue turn-interaction into account. The experimental results demonstrate MIE is a promising solution to extract medical information from doctor-patient dialogues.

Named Entity Recognition as Dependency Parsing [Website][PDF]
Juntao Yu, Bernd Bohnet, and Massimo Poesio 6:00–7:00

Named Entity Recognition (NER) is a fundamental task in Natural Language Processing, concerned with identifying spans of text expressing references to entities. NER research is often focused on flat entities only (flat NER), ignoring the fact that entity references can be nested, as in [Bank of [China]] (Finkel and Manning, 2009). In this paper, we use ideas from graph-based dependency parsing to provide our model a global view on the input via a biaffine model (Dozat and Manning, 2017). The biaffine model scores pairs of start and end tokens in a sentence which we use to explore all spans, so that the model is able to predict named entities accurately. We show that the model works well for both nested and flat NER through evaluation on 8 corpora and achieving SoTA performance on all of them, with accuracy gains of up to 2.2 percentage points.

Neighborhood Matching Network for Entity Alignment [Website][PDF]
Yuting Wu, Xiao Liu, Yansong Feng, Zheng Wang, and Dongyan Zhao 6:00–7:00

Structural heterogeneity between knowledge graphs is an outstanding challenge for entity alignment. This paper presents Neighborhood Matching Network (NMN), a novel entity alignment framework for tackling the structural heterogeneity challenge. NMN estimates the similarities between entities to capture both the topological structure and the neighborhood difference. It provides two innovative components for better learning representations for entity alignment. It first uses a novel graph sampling method to distill a discriminative neighborhood for each entity. It then adopts a cross-graph neighborhood matching module to jointly encode the neighborhood difference for a given entity pair. Such strategies allow NMN to effectively construct matching-oriented entity representations while ignoring noisy neighbors that have a negative impact on the alignment task. Extensive experiments performed on three entity alignment datasets show that NMN can well estimate the neighborhood similarity in more tough cases and significantly outperforms 12 previous state-of-the-art methods.

Relation Extraction with Explanation [Website][PDF]
Hamed Shahbazi, Xiaoli Fern, Reza Ghaeini, and Prasad Tadepalli 6:00–7:00

Recent neural models for relation extraction with distant supervision alleviate the impact of irrelevant sentences in a bag by learning importance weights for the sentences. Efforts thus far have focused on improving extraction accuracy but little is known about their explanability. In this work we annotate a test set with ground-truth sentence-level explanations to evaluate the quality of explanations afforded by the relation extraction models. We demonstrate that replacing the entity mentions in the sentences with their fine-grained entity types not only enhances extraction accuracy but also improves explanation. We also propose to automatically generate "distractor" sentences to augment the bags and train the model to ignore the distractors. Evaluations on the widely used FB-NYT dataset show that our methods achieve new state-of-the-art accuracy while improving model explanability.

Representation Learning for Information Extraction from Form-like Documents [Website][PDF]
Bodhisattwa Prasad Majumder, Navneet Potti, Sandeep Tata, James Bradley Wendt, Qi Zhao, and Marc Najork 6:00–7:00

We propose a novel approach using representation learning for tackling the problem of extracting structured information from form-like document images. We propose an extraction system that uses knowledge of the types of the target fields to generate extraction candidates and a neural network architecture that learns a dense representation of each candidate based on neighboring words in the document. These learned representations are not only useful in solving the extraction task for unseen document templates from two different domains but are also interpretable, as we show using loss cases.

Single-/Multi-Source Cross-Lingual NER via Teacher-Student Learning on Unlabeled Data in Target Language [Website] [PDF]
Qianhui Wu, Zijia Lin, Börje Karlsson, Jian-Guang LOU, and Biqing Huang 6:00–7:00

To better tackle the named entity recognition (NER) problem on languages with little/no labeled data, cross-lingual NER must effectively leverage knowledge learned from source languages with rich labeled data. Previous works on cross-lingual NER are mostly based on label projection with pairwise texts or direct model transfer. However, such methods either are not applicable if the labeled data in the source languages is unavailable, or do not leverage information contained in unlabeled data in the target language. In this paper, we propose a teacher-student learning method to address such limitations, where NER models in the source languages are used as teachers to train a student model on unlabeled data in the target language. The proposed method works for both single-source and multi-source cross-lingual NER. For the latter, we further propose a similarity measuring method to better weight the supervision from different teacher models. Extensive experiments for 3 target languages on benchmark datasets well demonstrate that our method outperforms existing state-of-the-art methods for both single-source and multi-source cross-lingual NER.

Synchronous Double-channel Recurrent Network for Aspect-Opinion Pair Extraction [Website] [PDF]
Shaowei Chen, Jie Liu, Yu Wang, Wenzheng Zhang, and Ziming Chi 6:00–7:00

Opinion entity extraction is a fundamental task in fine-grained opinion mining. Related studies generally extract aspects and/or opinion expressions without recognizing the relations between them. However, the relations are crucial for downstream tasks, including sentiment classification, opinion summarization, etc. In this paper, we explore Aspect-Opinion Pair Extraction (AOPE) task, which aims at extracting aspects and opinion expressions in pairs. To deal with this task, we propose Synchronous Double-channel Recurrent Network (SDRN) mainly consisting of an opinion entity extraction unit, a relation detection unit, and a synchronization unit. The opinion entity extraction unit and the relation detection unit are developed as two channels to extract opinion entities and relations simultaneously. Furthermore, within the synchronization unit, we design Entity Synchronization Mechanism (ESM) and Relation Synchronization Mechanism (RSM) to enhance the mutual benefit on the above two channels. To verify the performance of SDRN, we manually build three datasets based on SemEval 2014 and 2015 benchmarks. Extensive experiments demonstrate that SDRN achieves state-of-the-art performances.

Session 11B: Language Grounding to Vision, Robotics and Beyond-5

Cross-modal Coherence Modeling for Caption Generation [Website][PDF]
Malihe Alikhani, Piyush Sharma, Shengjie Li, Radu Soricut, and Matthew Stone 6:00–7:00

We use coherence relations inspired by computational models of discourse to study the information needs and goals of image captioning. Using an annotation protocol specifically devised for capturing image–caption coherence relations, we annotate 10,000 instances from publicly-available image–caption pairs. We introduce a new task for learning inferences in imagery and text, coherence relation prediction, and show that these coherence annotations can be exploited to learn relation classifiers as an intermediary step, and also train coherence-aware, controllable image captioning models. The results show a dramatic improvement in the consistency and quality of the generated captions with respect to information needs specified via coherence relations.

Knowledge Supports Visual Language Grounding: A Case Study on Colour Terms [Website][PDF]
Simeon Schüz and Sina Zarrieß 6:00–7:00

In human cognition, world knowledge supports the perception of object colours: knowing that trees are typically green helps to perceive their colour in certain contexts. We go beyond previous studies on colour terms using isolated colour swatches and study visual grounding of colour terms in realistic objects. Our models integrate processing of visual information and object-specific knowledge via hard-coded (late) or learned (early) fusion. We find that both models consistently outperform a bottom-up baseline that predicts colour terms solely from visual inputs, but show interesting differences when predicting atypical colours of so-called colour diagnostic objects. Our models also achieve promising results when tested on new object categories not seen during training.

Span-based Localizing Network for Natural Language Video Localization [Website][PDF]
Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou 6:00–7:00

Given an untrimmed video and a text query, natural language video localization (NLVL) is to locate a matching span from the video that semantically corresponds to the query. Existing solutions formulate NLVL either as a ranking task and apply multimodal matching architecture, or as a regression task to directly regress the target video span. In this work, we address NLVL task with a span-based QA approach by treating the input video as text passage. We propose a video span localizing network (VSLNet), on top of the standard span-based QA framework, to address NLVL. The proposed VSLNet tackles the differences between NLVL and span-based QA through a simple and yet effective query-guided highlighting (QGH) strategy. The QGH guides VSLNet to search for matching video span within a highlighted region. Through extensive experiments on three benchmark datasets, we show that the proposed VSLNet outperforms the state-of-the-art methods; and adopting span-based QA framework is a promising direction to solve NLVL.

Words Aren't Enough, Their Order Matters: On the Robustness of Grounding Visual Referring Expressions [Website][PDF]
Arjun Akula, Spandana Gella, Yaser Al-Onaizan, Song-Chun Zhu, and Siva Reddy 6:00–7:00

Visual referring expression recognition is a challenging task that requires natural language understanding in the context of an image. We critically examine RefCOCOg, a standard benchmark for this task, using a human study and show that 83.7% of test instances do not require reasoning on linguistic structure, i.e., words are enough to identify the target object, the word order doesn't matter. To measure the true progress of existing models, we split the test set into two sets, one which requires reasoning on linguistic structure and the other which doesn't. Additionally, we create an out-of-distribution dataset Ref-Adv by asking crowdworkers to perturb in-domain examples such that the target object changes. Using these datasets, we empirically show that existing methods fail to exploit linguistic structure and are 12% to 23% lower in performance than the established progress for this task. We also propose two methods, one based on contrastive learning and the other based on multi-task learning, to increase the robustness of ViLBERT, the current state-of-the-art model for this task. Our datasets are publicly available at https://github.com/aws/aws-refcocog-adv.

Session 11B: Machine Learning for NLP-12

A Mixture of h - 1 Heads is Better than h Heads [Website][PDF]
Hao Peng, Roy Schwartz, Dianqi Li, and Noah A. Smith 6:00–7:00

Multi-head attentive neural architectures have achieved state-of-the-art results on a variety of natural language processing tasks. Evidence has shown that they are overparameterized; attention heads can be pruned without significant performance loss. In this work, we instead "reallocate" them—the model learns to activate different heads on different inputs. Drawing connections between multi-head attention and mixture of experts, we propose the mixture of attentive experts model (MAE). MAE is trained using a block coordinate descent algorithm that alternates between updating (1) the responsibilities of the experts and (2) their parameters. Experiments on machine translation and language modeling show that MAE outperforms strong baselines on both tasks. Particularly, on the WMT14 English to German translation dataset, MAE improves over "transformer-base" by 0.8 BLEU, with a comparable number of parameters. Our analysis shows that our model learns to specialize different experts to different inputs.

Dependency Graph Enhanced Dual-transformer Structure for Aspect-based Sentiment Classification [Website][PDF]
Hao Tang, Donghong Ji, Chenliang Li, and Qiji Zhou 6:00–7:00

Aspect-based sentiment classification is a popular task aimed at identifying the corresponding emotion of a specific aspect. One sentence may contain various sentiments for different aspects. Many sophisticated methods such as attention mechanism and Convolutional Neural Networks (CNN) have been widely employed for handling this challenge. Recently, semantic dependency tree implemented by Graph Convolutional Networks (GCN) is introduced to describe the inner connection between aspects and the associated emotion words. But the improvement is limited due to the noise and instability of dependency trees. To this end, we propose a dependency graph enhanced dual-transformer network (named DGEDT) by jointly considering the flat representations learnt from Transformer and graph-based representations learnt from the corresponding dependency graph in an iterative interaction manner. Specifically, a dual-transformer structure is devised in DGEDT to support mutual reinforcement between the flat representation learning and graph-based representation learning. The idea is to allow the dependency graph to guide the representation learning of the transformer encoder and vice versa. The results on five datasets demonstrate that the proposed DGEDT outperforms all state-of-the-art alternatives with a large margin.

Differentiable Window for Dynamic Local Attention [Website][PDF]
Thanh-Tung Nguyen, Xuan-Phi Nguyen, Shafiq Joty, and Xiaoli Li 6:00–7:00

We propose Differentiable Window, a new neural module and general purpose component for dynamic window selection. While universally applicable, we demonstrate a compelling use case of utilizing Differentiable Window to improve standard attention modules by enabling more focused attentions over the input regions. We propose two variants of Differentiable Window, and integrate them within the Transformer architecture in two novel ways. We evaluate our proposed approach on a myriad of NLP tasks, including machine translation, sentiment analysis, subject-verb agreement and language modeling. Our experimental results demonstrate consistent and sizable improvements across all tasks.

Evaluating and Enhancing the Robustness of Neural Network-based Dependency Parsing Models with Adversarial Examples [Website][PDF]
Xiaoqing Zheng, Jiehang Zeng, Yi Zhou, Cho-Jui Hsieh, Minhao Cheng, and Xuanjing Huang 6:00–7:00

Despite achieving prominent performance on many important tasks, it has been reported that neural networks are vulnerable to adversarial examples. Previously studies along this line mainly focused on semantic tasks such as sentiment analysis, question answering and reading comprehension. In this study, we show that adversarial examples also exist in dependency parsing: we propose two approaches to study where and how parsers make mistakes by searching over perturbations to existing texts at sentence and phrase levels, and design algorithms to construct such examples in both of the black-box and white-box settings. Our experiments with one of state-of-the-art parsers on the English Penn Treebank (PTB) show that up to 77% of input examples admit adversarial perturbations, and we also show that the robustness of parsing models can be improved by crafting high-quality adversaries and including them in the training stage, while suffering little to no performance drop on the clean input data.

Exploiting Syntactic Structure for Better Language Modeling: A Syntactic Distance Approach [Website][PDF]
Wenyu Du, Zhouhan Lin, Yikang Shen, Timothy J. O'Donnell, Yoshua Bengio, and Yue Zhang 6:00–7:00

It is commonly believed that knowledge of syntactic structure should improve language modeling. However, effectively and computationally efficiently incorporating syntactic structure into neural language models has been a challenging topic. In this paper, we make use of a multi-task objective, i.e., the models simultaneously predict words as well as ground truth parse trees in a form called "syntactic distances", where information between these two separate objectives shares the same intermediate representation. Experimental results on the Penn Treebank and Chinese Treebank datasets show that when ground truth parse trees are provided as additional training signals, the model is able to achieve lower perplexity and induce trees with better quality.

Learning Architectures from an Extended Search Space for Language Modeling [Website][PDF]
Yinqiao Li, Chi Hu, Yuhao Zhang, Nuo Xu, Yufan Jiang, Tong Xiao, Jingbo Zhu, Tongran Liu, and changliang li changliang 6:00–7:00

Neural architecture search (NAS) has advanced significantly in recent years but most NAS systems restrict search to learning architectures of a recurrent or convolutional cell. In this paper, we extend the search space of NAS. In particular, we present a general approach to learn both intra-cell and inter-cell architectures (call it ESS). For a better search result, we design a joint learning method to perform intra-cell and inter-cell NAS simultaneously. We implement our model in a differentiable architecture search system. For recurrent neural language modeling, it outperforms a strong baseline significantly on the PTB and WikiText data, with a new state-of-the-art on PTB. Moreover, the learned architectures show good transferability to other systems. E.g., they improve state-of-the-art systems on the CoNLL and WNUT named entity recognition (NER) tasks and CoNLL chunking task, indicating a promising line of research on large-scale pre-learned architectures.

The Right Tool for the Job: Matching Model and Instance Complexities [Website] [PDF]
Roy Schwartz, Gabriel Stanovsky, Swabha Swayamdipta, Jesse Dodge, and Noah A. Smith 6:00–7:00

As NLP models become larger, executing a trained model requires significant computational resources incurring monetary and environmental costs. To better respect a given inference budget, we propose a modification to contextual representation fine-tuning which, during inference, allows for an early (and fast) "exit" from neural network calculations for simple instances, and late (and accurate) exit for hard instances. To achieve this, we add classifiers to different layers of BERT and use their calibrated confidence scores to make early exit decisions. We test our proposed modification on five different datasets in two tasks: three text classification datasets and two natural language inference benchmarks. Our method presents a favorable speed/accuracy tradeoff in almost all cases, producing models which are up to five times faster than the state of the art, while preserving their accuracy. Our method also requires almost no additional training resources (in either time or parameters) compared to the baseline BERT model. Finally, our method alleviates the need for costly retraining of multiple models at different levels of efficiency; we allow users to control the inference speed/accuracy tradeoff using a single trained model, by setting a single variable at inference time. We publicly release our code.

Session 11B: Phonology, Morphology and Word Segmentation-3

Bootstrapping Techniques for Polysynthetic Morphological Analysis [Website][PDF]
William Lane and Steven Bird 6:00–7:00

Polysynthetic languages have exceptionally large and sparse vocabularies, thanks to the number of morpheme slots and combinations in a word. This complexity, together with a general scarcity of written data, poses a challenge to the development of natural language technologies. To address this challenge, we offer linguistically-informed approaches for bootstrapping a neural morphological analyzer, and demonstrate its application to Kunwinjku, a polysynthetic Australian language. We generate data from a finite state transducer to train an encoder-decoder model. We improve the model by "hallucinating" missing linguistic structure into the training data, and by resampling from a Zipf distribution to simulate a more natural distribution of morphemes. The best model accounts for all instances of reduplication in the test set and achieves an accuracy of 94.7% overall, a 10 percentage point improvement over the FST baseline. This process demonstrates the feasibility of bootstrapping a neural morph analyzer from minimal resources.

Coupling Distant Annotation and Adversarial Training for Cross-Domain Chinese Word Segmentation [Website][PDF]
Ning Ding, Dingkun Long, Guangwei Xu, Muhua Zhu, Pengjun Xie, Xiaobin Wang, and Haitao Zheng
6:00–7:00

Fully supervised neural approaches have achieved significant progress in the task of Chinese word segmentation (CWS). Nevertheless, the performance of supervised models always drops gravely if the domain shifts due to the distribution gap across domains and the out of vocabulary (OOV) problem. In order to simultaneously alleviate the issues, this paper intuitively couples distant annotation and adversarial training for cross-domain CWS. 1) We rethink the essence of "Chinese words" and design an automatic distant annotation mechanism, which does not need any supervision or pre-defined dictionaries on the target domain. The method could effectively explore domain-specific words and distantly annotate the raw texts for the target domain. 2) We further develop a sentence-level adversarial training procedure to perform noise reduction and maximum utilization of the source domain information. Experiments on multiple real-world datasets across various domains show the superiority and robustness of our model, significantly outperforming previous state-of-the-arts cross-domain CWS methods.

Modeling Morphological Typology for Unsupervised Learning of Language Morphology [Website][PDF]
Hongzhi Xu, Jordan Kodner, Mitchell Marcus, and Charles Yang 6:00–7:00

This paper describes a language-independent model for fully unsupervised morphological analysis that exploits a universal framework leveraging morphological typology. By modeling morphological processes including suffixation, prefixation, infixation, and full and partial reduplication with constrained stem change rules, our system effectively constrains the search space and offers a wide coverage in terms of morphological typology. The system is tested on nine typologically and genetically diverse languages, and shows superior performance over leading systems. We also investigate the effect of an oracle that provides only a handful of bits per language to signal morphological type.

Predicting Declension Class from Form and Meaning [Website][PDF]
Adina Williams, Tiago Pimentel, Hagen Blix, Arya D. McCarthy, Eleanor Chodroff, and Ryan Cotterell
6:00–7:00

The noun lexica of many natural languages are divided into several declension classes with characteristic morphological properties. Class membership is far from deterministic, but the phonological form of a noun and/or its meaning can often provide imperfect clues. Here, we investigate the strength of those clues. More specifically, we operationalize this by measuring how much information, in bits, we can glean about declension class from knowing the form and/or meaning of nouns. We know that form and meaning are often also indicative of grammatical gender—which, as we quantitatively verify, can itself share information with declension class—so we also control for gender. We find for two Indo-European languages (Czech and German) that form and meaning respectively share significant amounts of information with class (and contribute additional information above and beyond gender). The three-way interaction between class, form, and meaning (given gender) is also significant. Our study is important for two reasons: First, we introduce a new method that provides additional quantitative support for a classic linguistic finding that form and meaning are relevant for the classification of nouns into declensions. Secondly, we show not only that individual declensions classes vary in the strength of their clues within a language, but also that these variations themselves vary across languages.

Unsupervised Morphological Paradigm Completion [Website][PDF]
Huiming Jin, Liwei Cai, Yihui Peng, Chen Xia, Arya McCarthy, and Katharina Kann 6:00–7:00

We propose the task of unsupervised morphological paradigm completion. Given only raw text and a lemma list, the task consists of generating the morphological paradigms, i.e., all inflected forms, of the lemmas. From a natural language processing (NLP) perspective, this is a challenging unsupervised task, and high-performing systems have the potential to improve tools for low-resource languages or to assist linguistic annotators. From a cognitive science perspective, this can shed light on how children acquire morphological knowledge. We further introduce a system for the task, which generates morphological paradigms via the following steps: (i) EDIT TREE retrieval, (ii) additional lemma retrieval, (iii) paradigm size discovery, and (iv) inflection generation. We perform an evaluation on 14 typologically diverse languages. Our system outperforms trivial baselines with ease and, for some languages, even obtains a higher accuracy than minimally supervised systems.

Session 11B: Question Answering-9

Document Modeling with Graph Attention Networks for Multi-grained Machine Reading Comprehension [Website][PDF]
Bo Zheng, Haoyang Wen, Yaobo Liang, Nan Duan, Wanxiang Che, Daxin Jiang, Ming Zhou, and Ting Liu 6:00–7:00

Natural Questions is a new challenging machine reading comprehension benchmark with two-grained answers, which are a long answer (typically a paragraph) and a short answer (one or more entities inside the long answer). Despite the effectiveness of existing methods on this benchmark, they treat these two sub-tasks individually during training while ignoring their dependencies. To address this issue, we present a novel multi-grained machine reading comprehension framework that focuses on modeling documents at their hierarchical nature, which are different levels of granularity: documents, paragraphs, sentences, and tokens. We utilize graph attention networks to obtain different levels of representations so that they can be learned simultaneously. The long and short answers can be extracted from paragraph-level representation and token-level representation, respectively. In this way, we can model the dependencies between the two-grained answers to provide evidence for each other. We jointly train the two sub-tasks, and our experiments show that our approach significantly outperforms previous systems at both long and short answer criteria.

Harvesting and Refining Question-Answer Pairs for Unsupervised QA [Website][PDF]
Zhongli Li, Wenhui Wang, Li Dong, Furu Wei, and Ke Xu 6:00–7:00

Question Answering (QA) has shown great success thanks to the availability of large-scale datasets and the effectiveness of neural models. Recent research works have attempted to extend these successes to the settings with few or no labeled data available. In this work, we introduce two approaches to improve unsupervised QA. First, we harvest lexically and syntactically divergent questions from Wikipedia to automatically construct a corpus of question-answer pairs (named as RefQA). Second, we take advantage of the QA model to extract more appropriate answers, which iteratively refines data over RefQA. We conduct experiments on SQuAD 1.1, and NewsQA by fine-tuning BERT without access to manually annotated data. Our approach outperforms previous unsupervised approaches by a large margin, and is competitive with early supervised models. We also show the effectiveness of our approach in the few-shot learning setting.

Low-Resource Generation of Multi-hop Reasoning Questions [Website][PDF]
Jianxing Yu, Wei Liu, Shuang Qiu, Qinliang Su, Kai Wang, Xiaojun Quan, and Jian Yin 6:00–7:00

This paper focuses on generating multi-hop reasoning questions from the raw text in a low resource circumstance. Such questions have to be syntactically valid and need to logically correlate with the answers by deducing over multiple relations on several sentences in the text. Specifically, we first build a multi-hop generation model and guide it to satisfy the logical rationality by the reasoning chain extracted from a given text. Since the labeled data is limited and insufficient for training, we propose to learn the model with the help of a large scale of unlabeled data that is much easier to obtain. Such data contains rich expressive forms of the questions with structural patterns on syntax and semantics. These patterns can be estimated by the neural hidden semi-Markov model using latent variables. With latent patterns as a prior, we can regularize the generation model and produce the optimal results. Experimental results on the HotpotQA data set demonstrate the effectiveness of our model. Moreover, we apply the generated results to the task of machine reading comprehension and achieve significant performance improvements.

R4C: A Benchmark for Evaluating RC Systems to Get the Right Answer for the Right Reason [Website][PDF]
Naoya Inoue, Pontus Stenetorp, and Kentaro Inui 6:00–7:00

Recent studies have revealed that reading comprehension (RC) systems learn to exploit annotation artifacts and other biases in current datasets. This prevents the community from reliably measuring the progress of RC systems. To address this issue, we introduce R4C, a new task for evaluating RC systems' internal reasoning. R4C requires giving not only answers but also derivations: explanations that justify predicted answers. We present a reliable, crowdsourced framework for scalably annotating RC datasets with derivations. We create and publicly release the R4C dataset, the first, quality-assured dataset consisting of 4.6k questions, each of which is annotated with 3 reference derivations (i.e. 13.8k derivations). Experiments show that our automatic evaluation metrics using multiple reference derivations are reliable, and that R4C assesses different skills from an existing benchmark.

Recurrent Chunking Mechanisms for Long-Text Machine Reading Comprehension [Website][PDF]
Hongyu Gong, Yelong Shen, Dian Yu, Jianshu Chen, and Dong Yu 6:00–7:00

In this paper, we study machine reading comprehension (MRC) on long texts: where a model takes as inputs a lengthy document and a query, extracts a text span from the document as an answer. State-of-the-art models (e.g., BERT) tend to use a stack of transformer layers that are pre-trained from a large number of unlabeled language corpora to encode the joint contextual information of query and document. However, these transformer models can only take as input a fixed-length (e.g., 512) text. To deal with even longer text inputs, previous approaches usually chunk them into *equally-spaced* segments and predict answers based on each segment independently without considering the information from other segments. As a result, they may form segments that fail to cover complete answers or retain insufficient contexts around the correct answer required for question answering. Moreover, they are less capable of answering questions that need cross-segment information. We propose to let a model learn to chunk in a more flexible way via reinforcement learning: a model can decide the next segment that it wants to process in either direction. We also apply recurrent mechanisms to enable information to flow across segments. Experiments on three MRC tasks – CoQA, QuAC, and TriviaQA – demonstrate the effectiveness of our proposed recurrent chunking mechanisms: we can

obtain segments that are more likely to contain complete answers and at the same time provide sufficient contexts around the ground truth answers for better predictions.

RikiNet: Reading Wikipedia Pages for Natural Question Answering [Website][PDF]
Dayiheng Liu, Yeyun Gong, Jie Fu, Yu Yan, Jiusheng Chen, Daxin Jiang, Jiancheng Lv, and Nan Duan
6:00–7:00

Reading long documents to answer open-domain questions remains challenging in natural language understanding. In this paper, we introduce a new model, called RikiNet, which reads Wikipedia pages for natural question answering. RikiNet contains a dynamic paragraph dual-attention reader and a multi-level cascaded answer predictor. The reader dynamically represents the document and question by utilizing a set of complementary attention mechanisms. The representations are then fed into the predictor to obtain the span of the short answer, the paragraph of the long answer, and the answer type in a cascaded manner. On the Natural Questions (NQ) dataset, a single RikiNet achieves 74.3 F1 and 57.9 F1 on long-answer and short-answer tasks. To our best knowledge, it is the first single model that outperforms the single human performance. Furthermore, an ensemble RikiNet obtains 76.1 F1 and 61.3 F1 on long-answer and short-answer tasks, achieving the best performance on the official NQ leaderboard.

Session 11B Semantics: Sentence Level-6

[TACL] AMR-To-Text Generation with Graph Transformer [Website][PDF]
Tianming Wang, Xiaojun Wan, and Hanqi Jin 6:00–7:00

Abstract meaning representation (AMR)-to-text generation is the challenging task of generating natural language texts from AMR graphs, where nodes represent concepts and edges denote relations. The current state-of-the-art methods use graph-to-sequence models; however, they still cannot significantly outperform the previous sequence-to-sequence models or statistical approaches. In this paper, we propose a novel graph-to-sequence model (Graph Transformer) to address the above-mentioned task. The model directly encodes the AMR graphs and learns the node representations. A pairwise interaction function is used for computing the semantic relations between the concepts. Moreover, attention mechanisms are employed for aggregating the information from the incoming and outgoing neighbors, which help the model to capture the semantic information effectively. Our model outperforms the state-of-the-art neural approach by 1.5 BLEU points on LDC2015E86 and 4.8 BLEU points on LDC2017T10 and achieves new state-of-the-art performances.

Parsing into Variable-in-situ Logico-Semantic Graphs [Website][PDF]
Yufei Chen and Weiwei Sun 6:00–7:00

We propose variable-in-situ logico-semantic graphs to bridge the gap between semantic graph and logical form parsing. The new type of graph-based meaning representation allows us to include analysis for scope-related phenomena, such as quantification, negation and modality, in a way that is consistent with the state-of-the-art underspecification approach. Moreover, the well-formedness of such a graph is clear, since model-theoretic interpretation is available. We demonstrate the effectiveness of this new perspective by developing a new state-of-the-art semantic parser for English Resource Semantics. At the core of this parser is a novel neural graph rewriting system which combines the strengths of Hyperedge Replacement Grammar, a knowledge-intensive model, and Graph Neural Networks, a data-intensive model. Our parser achieves an accuracy of 92.39% in terms of elementary dependency match, which is a 2.88 point improvement over the best data-driven model in the literature. The output of our parser is highly coherent: at least 91% graphs are valid, in that they allow at least one sound scope-resolved logical form.

Semantic Parsing for English as a Second Language [Website][PDF]
Yuanyuan Zhao, Weiwei Sun, junjie cao junjie, and Xiaojun Wan 6:00–7:00

This paper is concerned with semantic parsing for English as a second language (ESL). Motivated by the theoretical emphasis on the learning challenges that occur at the syntax-semantics interface during second language acquisition, we formulate the task based on the divergence between literal and intended meanings. We combine the complementary strengths of English Resource Grammar, a linguistically-precise hand-crafted deep grammar, and TLE, an existing manually annotated ESL UD-TreeBank with a novel reranking model. Experiments demonstrate that in comparison to human annotations, our method can obtain a very promising SemBanking quality. By means of the newly created corpus, we evaluate state-of-the-art semantic parsing as well as grammatical error correction models. The evaluation profiles the performance of neural NLP techniques for handling ESL data and suggests some research directions.

Semi-Supervised Semantic Dependency Parsing Using CRF Autoencoders [Website][PDF]
Zixia Jia, Youmi Ma, Jiong Cai, and Kewei Tu 6:00–7:00

Semantic dependency parsing, which aims to find rich bi-lexical relationships, allows words to have multiple dependency heads, resulting in graph-structured representations. We propose an approach to semi-supervised learning of semantic dependency parsers based on the CRF autoencoder framework. Our encoder is a discriminative neural semantic dependency parser that predicts the latent parse graph of the input sentence. Our decoder is a generative neural model that reconstructs the input sentence conditioned on the latent parse graph. Our model is arc-factored and therefore parsing and learning are both tractable. Experiments show our model achieves significant and consistent improvement over the supervised baseline.

Unsupervised Dual Paraphrasing for Two-stage Semantic Parsing [Website][PDF]
Ruisheng Cao, Su Zhu, Chenyu Yang, Chen Liu, Rao Ma, Yanbin Zhao, Lu Chen, and Kai Yu 6:00–7:00

One daunting problem for semantic parsing is the scarcity of annotation. Aiming to reduce nontrivial human labor, we propose a two-stage semantic parsing framework, where the first stage utilizes an unsupervised paraphrase model to convert an unlabeled natural language utterance into the canonical utterance. The downstream naive semantic parser accepts the intermediate output and returns the target logical form. Furthermore, the entire training process is split into two phases: pre-training and cycle learning. Three tailored self-supervised tasks are introduced throughout training to activate the unsupervised paraphrase model. Experimental results on benchmarks Overnight and GeoGranno demonstrate that our framework is effective and compatible with supervised training.

Session 11B: Student Research Workshop

Feature Difference Makes Sense: A medical image captioning model exploiting feature difference and tag information [Website][PDF]
Hyeryun Park, Kyungmo Kim, Jooyoung Yoon, Seongkeun Park, and Jinwook Choi 6:00–7:00

Medical image captioning can reduce the workload of physicians and save time and expense by automatically generating reports. However, current datasets are small and limited, creating additional challenges for researchers. In this study, we propose a feature difference and tag information combined long short-term memory (LSTM) model for chest x-ray report generation. A feature vector extracted from the image conveys visual information, but its ability to describe the image is limited. Other image captioning studies exhibited improved performance by exploiting feature differences, so the proposed model also utilizes them. First, we propose a difference and tag (DiTag) model containing the difference between the patient and normal images. Then, we propose a multi-difference and tag (mDiTag) model that also contains information about low-level differences, such as contrast, texture, and localized area. Evaluation of the proposed models demonstrates that the mDiTag model provides more information to generate captions and outperforms all other models.

Multi-Task Neural Model for Agglutinative Language Translation [Website][PDF]
Yirong Pan, Xiao Li, Yating Yang, and Rui Dong 6:00–7:00

Neural machine translation (NMT) has achieved impressive performance recently by using large-scale parallel corpora. However, it struggles in the low-resource and morphologically-rich scenarios of agglutinative language translation task. Inspired by the finding that monolingual data can greatly improve the NMT performance, we propose a multi-task neural model that jointly learns to perform bi-directional translation and agglutinative language stemming. Our approach employs the shared encoder and decoder to train a single model without changing the standard NMT architecture but instead adding a token before each source-side sentence to specify the desired target outputs of the two different tasks. Experimental results on Turkish-English and Uyghur-Chinese show that our proposed approach can significantly improve the translation performance on agglutinative languages by using a small amount of monolingual data.

Demo Session 1C

Time: *6:30–7:15*

Penman: An Open-Source Library and Tool for AMR Graphs [Website][PDF]
Michael Wayne Goodman

Abstract Meaning Representation (AMR) (Banarescu et al., 2013) is a framework for semantic dependencies that encodes its rooted and directed acyclic graphs in a format called PENMAN notation. The format is simple enough that users of AMR data often write small scripts or libraries for parsing it into an internal graph representation, but there is enough complexity that these users could benefit from a more sophisticated and well-tested solution. The open-source Python library Penman provides a robust parser, functions for graph inspection and manipulation, and functions for formatting graphs into PENMAN notation. Many functions are also available in a command-line tool, thus extending its utility to non-Python setups.

Demo Session 2A

Time: *8:00–8:45*

Embedding-based Scientific Literature Discovery in a Text Editor Application [Website] [PDF]
Onur Gökçe, Jonathan Prada, Nikola I. Nikolov, Nianlong Gu, and Richard H.R. Hahnloser

Each claim in a research paper requires all relevant prior knowledge to be discovered, assimilated, and appropriately cited. However, despite the availability of powerful search engines and sophisticated text editing software, discovering relevant papers and integrating the knowledge into a manuscript remain complex tasks associated with high cognitive load. To define comprehensive search queries requires strong motivation from authors, irrespective of their familiarity with the research field. Moreover, switching between independent applications for literature discovery, bibliography management, reading papers, and writing text burdens authors further and interrupts their creative process. Here, we present a web application that combines text editing and literature discovery in an interactive user interface. The application is equipped with a search engine that couples Boolean keyword filtering with nearest neighbor search over text embeddings, providing a discovery experience tuned to an author's manuscript and his interests. Our application aims to take a step towards more enjoyable and effortless academic writing. The demo of the application (https://SciEditorDemo2020.herokuapp.com) and a short video tutorial (https://youtu.be/pkdVU60IcRc) are available online.

Session 12A Overview – Wednesday, July 8, 2020 8:00–9:00

Track A *Discourse and Pragmatics-7* Abstracts	A Top-down Neural Architecture towards Text-level Parsing of Discourse Rhetorical Structure *Zhang, Xing, Kong, Li, and Zhou* [Website][PDF]	DRTS Parsing with Structure-Aware Encoding and Decoding *Fu, Zhang, Liu, and Zhang* [Website][PDF]	[TACL] Unsupervised Discourse Constituency Parsing Using Viterbi EM *Nishida and Nakayama* [Website][PDF]		
Track B *Information Extraction-5* Abstracts	A Two-Stage Masked LM Method for Term Set Expansion *Kushilevitz, Markovitch, and Goldberg* [Website][PDF]	Amalgamation of protein sequence, structure and textual information for improving protein-protein interaction identification *Dutta and Saha* [Website][PDF]	FLAT: Chinese NER Using Flat-Lattice Transformer *Li, Yan, Qiu, and Huang* [Website][PDF]	IMoJIE: Iterative Memory-Based Joint Open Information Extraction *Kolluru, Aggarwal, Rathore, and Chakrabarti* [Website][PDF]	Improving Entity Linking through Semantic Reinforced Entity Embeddings *Hou, Wang, He, and Zhou* [Website][PDF]
	Improving Event Detection via Open-domain Trigger Knowledge *Tong, Xu, Wang, Cao, Hou, Li, and Xie* [Website][PDF]	Improving Low-Resource Named Entity Recognition using Joint Sentence and Token Labeling *Kruengkrai, Nguyen, Aljunied, and Bing* [Website][PDF]	MIE: A Medical Information Extractor towards Medical Dialogues *Zhang, Jiang, Zhang, Liu, Cao, Liu, Liu, and Zhao* [Website][PDF]	Multi-Cell Compositional LSTM for NER Domain Adaptation *Jia and Zhang* [Website][PDF]	Neighborhood Matching Network for Entity Alignment *Wu, Liu, Feng, Wang, and Zhao* [Website][PDF]
	Pyramid: A Layered Model for Nested Named Entity Recognition *WANG, Shou, Chen, and Chen* [Website][PDF]	ReInceptionE: Relation-Aware Inception Network with Joint Local-Global Structural Information for Knowledge Graph Embedding *Xie, Zhou, Liu, and Huang* [Website][PDF]	Relabel the Noise: Joint Extraction of Entities and Relations via Cooperative Multiagents *Chen, Li, Lei, and Shen* [Website][PDF]	Simplify the Usage of Lexicon in Chinese NER *Ma, Peng, Zhang, Wei, and Huang* [Website][PDF]	
Track C *Information Retrieval and Text Mining-6* Abstracts	Document Translation vs. Query Translation for Cross-Lingual Information Retrieval in the Medical Domain *Saleh and Pecina* [Website][PDF]	Learning Robust Models for e-Commerce Product Search *Nguyen, Rao, and Subbian* [Website][PDF]			
Track D *Machine Learning for NLP-13* Abstracts	Dependency Graph Enhanced Dual-transformer Structure for Aspect-based Sentiment Classification *Tang, Ji, Li, and Zhou* [Website][PDF]	Differentiable Window for Dynamic Local Attention *Nguyen, Nguyen, Joty, and Li* [Website][PDF]	Generalized Entropy Regularization or: There's Nothing Special about Label Smoothing *Meister, Salesky, and Cotterell* [Website][PDF]	Highway Transformer: Self-Gating Enhanced Self-Attentive Networks *Chai, Jin, and Hou* [Website][PDF]	Low-Dimensional Hyperbolic Knowledge Graph Embeddings *Chami, Wolf, Juan, Sala, Ravi, and Ré* [Website][PDF]

Track E *Machine Translation-14* Abstracts	AdvAug: Robust Adversarial Augmentation for Neural Machine Translation *Cheng, Jiang, Macherey, and Eisenstein* [Website][PDF]	Classification-Based Self-Learning for Weakly Supervised Bilingual Lexicon Induction *Karan, Vulić, Korhonen, and Glavaš* [Website][PDF]	Contextual Neural Machine Translation Improves Translation of Cataphoric Pronouns *Wong, Maruf, and Haffari* [Website][PDF]	Gender in Danger? Evaluating Speech Translation Technology on the MuST-SHE Corpus *Bentivogli, Savoldi, Negri, Di Gangi, Cattoni, and Turchi* [Website][PDF]	Improving Neural Machine Translation with Soft Template Prediction *Yang, Ma, Zhang, Li, and Zhou* [Website][PDF]
	Uncertainty-Aware Curriculum Learning for Neural Machine Translation *Zhou, Yang, Wong, Wan, and Chao* [Website][PDF]	Worse WER, but Better BLEU? Leveraging Word Embedding as Intermediate in Multitask End-to-End Speech Translation *Chuang, Sung, Liu, and Lee* [Website][PDF]			
Track F *NLP Applications-9* Abstracts	Closing the Gap: Joint De-Identification and Concept Extraction in the Clinical Domain *Lange, Adel, and Strötgen* [Website][PDF]	CorefQA: Coreference Resolution as Query-based Span Prediction *Wu, Wang, Yuan, Wu, and Li* [Website][PDF]	Estimating predictive uncertainty for rumour verification models *Kochkina and Liakata* [Website][PDF]	From Zero to Hero: Human-In-The-Loop Entity Linking in Low Resource Domains *Klie, Eckart de Castilho, and Gurevych* [Website][PDF]	Language to Network: Conditional Parameter Adaptation with Natural Language Descriptions *Jin, Liu, Yan, Eichenberger, and Morency* [Website][PDF]
	Neural-DINF: A Neural Network based Framework for Measuring Document Influence *Tan, Yang, Li, Tang, Huang, and Zhuang* [Website][PDF]				
Track G *Sentence Level-7* Abstracts	[TACL] AMR-To-Text Generation with Graph Transformer *Wang, Wan, and Jin* [Website][PDF]	Controlled Crowdsourcing for High-Quality QA-SRL Annotation *Roit, Klein, Stepanov, Mamou, Michael, Stanovsky, Zettlemoyer, and Dagan* [Website][PDF]	Cross-Lingual Semantic Role Labeling with High-Quality Translated Training Corpus *Fei, Zhang, and Ji* [Website][PDF]	Semantic Parsing for English as a Second Language *Zhao, Sun, and Wan* [Website][PDF]	Sentence Meta-Embeddings for Unsupervised Semantic Textual Similarity *Poerner, Waltinger, and Schütze* [Website][PDF]
	Transition-based Semantic Dependency Parsing with Pointer Networks *Fernández-González and Gómez-Rodríguez* [Website][PDF]	Unsupervised Dual Paraphrasing for Two-stage Semantic Parsing *Cao, Zhu, Yang, Liu, Ma, Zhao, Chen, and Yu* [Website][PDF]	Word-level Textual Adversarial Attacking as Combinatorial Optimization *Zang, Qi, Yang, Liu, Zhang, Liu, and Sun* [Website][PDF]	tBERT: Topic Models and BERT Joining Forces for Semantic Similarity Detection *Peinelt, Nguyen, and Liakata* [Website][PDF]	

Track H *Sentiment Analysis, Stylistic Analysis, and Argument Mining-10* Abstracts	Conditional Augmentation for Aspect Term Extraction via Masked Sequence-to-Sequence Generation *Li, Chen, Quan, Ling, and Song* [Website][PDF]	Exploiting Personal Characteristics of Debaters for Predicting Persuasiveness *Al Khatib, Völske, Syed, Kolyada, and Stein* [Website][PDF]	Out of the Echo Chamber: Detecting Countering Debate Speeches *Orbach, Bilu, Toledo, Lahav, Jacovi, Aharonov, and Slonim* [Website][PDF]		
Track I *Student Research Workshop* Abstracts	Zero-shot North Korean to English Neural Machine Translation by Character Tokenization and Phoneme Decomposition *Kim, Hirasawa, and Komachi* [Website][PDF]	Research on Task Discovery for Transfer Learning in Deep Neural Networks *Akdemir* [Website][PDF]	uBLEU: Uncertainty-Aware Automatic Evaluation Method for Open-Domain Dialogue Systems *Yuma, Yoshinaga, and Toyoda* [Website][PDF]		
Track J *Summarization-7* Abstracts	Automatic Generation of Citation Texts in Scholarly Papers: A Pilot Study *Xing, Fan, and Wan* [Website][PDF]	Composing Elementary Discourse Units in Abstractive Summarization *Li, Wu, and Li* [Website][PDF]	Extractive Summarization as Text Matching *Zhong, Liu, Chen, Wang, Qiu, and Huang* [Website][PDF]	Heterogeneous Graph Neural Networks for Extractive Document Summarization *Wang, Liu, Zheng, Qiu, and Huang* [Website][PDF]	Jointly Learning to Align and Summarize for Neural Cross-Lingual Summarization *Cao, Liu, and Wan* [Website][PDF]
	Leveraging Graph to Improve Abstractive Multi-Document Summarization *Li, Xiao, Liu, Wu, Wang, and Du* [Website][PDF]	Multi-Granularity Interaction Network for Extractive and Abstractive Multi-Document Summarization *Jin, Wang, and Wan* [Website][PDF]			

Session 12A Details

Session 12A: Discourse and Pragmatics-7

A Top-down Neural Architecture towards Text-level Parsing of Discourse Rhetorical Structure [Website][PDF]
Longyin Zhang, Yuqing Xing, Fang Kong, Peifeng Li, and Guodong Zhou 8:00–9:00

Due to its great importance in deep natural language understanding and various down-stream applications, text-level parsing of discourse rhetorical structure (DRS) has been drawing more and more attention in recent years. However, all the previous studies on text-level discourse parsing adopt bottom-up approaches, which much limit the DRS determination on local information and fail to well benefit from global information of the overall discourse. In this paper, we justify from both computational and perceptive points-of-view that the top-down architecture is more suitable for text-level DRS parsing. On the basis, we propose a top-down neural architecture toward text-level DRS parsing. In particular, we cast discourse parsing as a recursive split point ranking task, where a split point is classified to different levels according to its rank and the elementary discourse units (EDUs) associated with it are arranged accordingly. In this way, we can determine the complete DRS as a hierarchical tree structure via an encoder-decoder with an internal stack. Experimentation on both the English RST-DT corpus and the Chinese CDTB corpus shows the great effectiveness of our proposed top-down approach towards text-level DRS parsing.

DRTS Parsing with Structure-Aware Encoding and Decoding [Website][PDF]
Qiankun Fu, Yue Zhang, Jiangming Liu, and Meishan Zhang 8:00–9:00

Discourse representation tree structure (DRTS) parsing is a novel semantic parsing task which has been concerned most recently. State-of-the-art performance can be achieved by a neural sequence-to-sequence model, treating the tree construction as an incremental sequence generation problem. Structural information such as input syntax and the intermediate skeleton of the partial output has been ignored in the model, which could be potentially useful for the DRTS parsing. In this work, we propose a structural-aware model at both the encoder and decoder phase to integrate the structural information, where graph attention network (GAT) is exploited for effectively modeling. Experimental results on a benchmark dataset show that our proposed model is effective and can obtain the best performance in the literature.

[TACL] Unsupervised Discourse Constituency Parsing Using Viterbi EM [Website][PDF]
Noriki Nishida and Hideki Nakayama 8:00–9:00

In this paper, we introduce an unsupervised discourse constituency parsing algorithm. We use Viterbi EM with a margin-based criterion to train a span-based discourse parser in an unsupervised manner. We also propose initialization methods for Viterbi training of discourse constituents based on our prior knowledge of text structures. Experimental results demonstrate that our unsupervised parser achieves comparable or even superior performance to fully supervised parsers. We also investigate discourse constituents that are learned by our method.

Session 12A: Information Extraction-5

A Two-Stage Masked LM Method for Term Set Expansion [Website][PDF]
Guy Kushilevitz, Shaul Markovitch, and Yoav Goldberg 8:00–9:00

We tackle the task of Term Set Expansion (TSE): given a small seed set of example terms from a semantic class, finding more members of that class. The task is of great practical utility, and also of theoretical utility as it requires generalization from few examples. Previous approaches to the TSE task can be characterized as either distributional or pattern-based. We harness the power of neural masked language models (MLM) and propose a novel TSE algorithm, which combines the pattern-based and distributional approaches. Due to the small size of the seed set, fine-tuning methods are not effective, calling for more creative use of the MLM. The gist of the idea is to use the MLM to first mine for informative patterns with respect to the seed set, and then to obtain more members of the seed class by generalizing these patterns. Our method outperforms state-of-the-art TSE algorithms. Implementation is available at: https://github.com/ guykush/TermSetExpansion-MPB/

Amalgamation of protein sequence, structure and textual information for improving protein-protein interaction identification [Website][PDF]
pratik Dutta and Sriparna Saha 8:00–9:00

An in-depth exploration of protein-protein interactions (PPI) is essential to understand the metabolism in addition to the regulations of biological entities like proteins, carbohydrates, and many more. Most of the recent PPI tasks in BioNLP domain have been carried out solely using textual data. In this paper, we argue that incorporating multimodal cues can improve the automatic identification of PPI. As a first step towards enabling the development of multimodal approaches for PPI identification, we have developed two multi-modal datasets which are extensions and multi-modal versions of two popular benchmark PPI corpora (BioInfer and HRPD50). Besides, existing textual modalities, two new modalities, 3D protein structure and underlying genomic sequence, are also added to each instance. Further, a novel deep multi-modal architecture is also implemented to efficiently predict the protein interactions from the developed datasets. A detailed experimental analysis reveals the superiority of the multi-modal approach in comparison to the strong baselines including unimodal approaches and state-of the-art methods over both the generated multi-modal datasets. The developed multi-modal datasets are available for use at https://github.com/sduttap16/MM_PPI_NLP.

FLAT: Chinese NER Using Flat-Lattice Transformer [Website][PDF]
Xiaonan Li, Hang Yan, Xipeng Qiu, and Xuanjing Huang 8:00–9:00

Recently, the character-word lattice structure has been proved to be effective for Chinese named entity recognition (NER) by incorporating the word information. However, since the lattice structure is complex and dynamic, the lattice-based models are hard to fully utilize the parallel computation of GPUs and usually have a low inference speed. In this paper, we propose FLAT: Flat-LAttice Transformer for Chinese NER, which converts the lattice structure into a flat structure consisting of spans. Each span corresponds to a character or latent word and its position in the original lattice. With the power of Transformer and well-designed position encoding, FLAT can fully leverage the lattice information and has an excellent parallel ability. Experiments on four datasets show FLAT outperforms other lexicon-based models in performance and efficiency.

IMoJIE: Iterative Memory-Based Joint Open Information Extraction [Website][PDF]
Keshav Kolluru, Samarth Aggarwal, Vipul Rathore, and Soumen Chakrabarti 8:00–9:00

While traditional systems for Open Information Extraction were statistical and rule-based, recently neural models have been introduced for the task. Our work builds upon CopyAttention, a sequence generation OpenIE model (Cui et. al. 18). Our analysis reveals that CopyAttention produces a constant number of extractions per sentence, and its extracted tuples often express redundant information. We present IMoJIE, an extension to CopyAttention, which produces the next extraction conditioned on all previously extracted tuples. This approach overcomes both shortcomings of CopyAttention, resulting in a variable number of diverse extractions per sentence. We train IMoJIE on training data bootstrapped from extractions of several non-neural systems, which have been automatically filtered to reduce redundancy and noise. IMoJIE outperforms CopyAttention by about 18 F1 pts, and a BERT-based strong baseline by 2 F1 pts, establishing a new state of the art for the task.

Improving Entity Linking through Semantic Reinforced Entity Embeddings [Website][PDF]
Feng Hou, Ruili Wang, Jun He, and Yi Zhou 8:00–9:00

Entity embeddings, which represent different aspects of each entity with a single vector like word embeddings, are a key component of neural entity linking models. Existing entity embeddings are learned from canonical Wikipedia articles and local contexts surrounding target entities. Such entity embeddings are effective, but too distinctive for linking models to learn contextual commonality. We propose a simple yet effective method, FGS2EE, to inject fine-grained semantic information into entity embeddings to reduce the distinctiveness and facilitate the learning of contextual commonality. FGS2EE first uses the embeddings of semantic type words to generate semantic embeddings, and then combines them with existing entity embeddings through linear aggregation. Extensive experiments show the effectiveness of such embeddings. Based on our entity embeddings, we achieved new sate-of-the-art performance on entity linking.

Improving Event Detection via Open-domain Trigger Knowledge [Website][PDF]
Meihan Tong, Bin Xu, Shuai Wang, Yixin Cao, Lei Hou, Juanzi Li, and Jun Xie 8:00–9:00

Event Detection (ED) is a fundamental task in automatically structuring texts. Due to the small scale of training data, previous methods perform poorly on unseen/sparsely labeled trigger words and are prone to overfitting densely labeled trigger words. To address the issue, we propose a novel Enrichment Knowledge Distillation (EKD) model to leverage external open-domain trigger knowledge to reduce the in-built biases to frequent trigger words in annotations. Experiments on benchmark ACE2005 show that our model outperforms nine strong baselines, is especially effective for unseen/sparsely labeled trigger words. The source code is released on https://github.com/shuaiwa16/ekd.git.

Improving Low-Resource Named Entity Recognition using Joint Sentence and Token Labeling [Website][PDF]
Canasai Kruengkrai, Thien Hai Nguyen, Sharifah Mahani Aljunied, and Lidong Bing 8:00–9:00

Exploiting sentence-level labels, which are easy to obtain, is one of the plausible methods to improve low-resource named entity recognition (NER), where token-level labels are costly to annotate. Current models for jointly learning sentence and token labeling are limited to binary classification. We present a joint model that supports multi-class classification and introduce a simple variant of self-attention that allows the model to learn scaling factors. Our model produces 3.78%, 4.20%, 2.08% improvements in F1 over the BiLSTM-CRF baseline on e-commerce product titles in three different low-resource languages: Vietnamese, Thai, and Indonesian, respectively.

MIE: A Medical Information Extractor towards Medical Dialogues [Website][PDF]
Yuanzhe Zhang, Zhongtao Jiang, Tao Zhang, Shiwan Liu, Jiarun Cao, Kang Liu, Shengping Liu, and Jun Zhao 8:00–9:00

Electronic Medical Records (EMRs) have become key components of modern medical care systems. Despite the merits of EMRs, many doctors suffer from writing them, which is time-consuming and tedious. We believe that automatically converting medical dialogues to EMRs can greatly reduce the burdens of doctors, and extracting information from medical dialogues is an essential step. To this end, we annotate online medical consultation dialogues in a window-sliding style, which is much easier than the sequential labeling annotation. We then propose a Medical Information Extractor (MIE) towards medical dialogues. MIE is able to extract mentioned symptoms, surgeries, tests, other information and their corresponding status. To tackle the particular challenges of the task, MIE uses a deep matching architecture, taking dialogue turn-interaction into account. The experimental results demonstrate MIE is a promising solution to extract medical information from doctor-patient dialogues.

Multi-Cell Compositional LSTM for NER Domain Adaptation [Website][PDF]
Chen Jia and Yue Zhang 8:00–9:00

Cross-domain NER is a challenging yet practical problem. Entity mentions can be highly different across domains. However, the correlations between entity types can be relatively more stable across domains. We investigate a multi-cell compositional LSTM structure for multi-task learning, modeling each entity type using a separate cell state. With the help of entity typed units, cross-domain knowledge transfer can be made in an entity type level. Theoretically, the resulting distinct feature distributions for each entity type make it more powerful for cross-domain transfer. Empirically, experiments on four few-shot and zero-shot datasets show our method significantly outperforms a series of multi-task learning methods and achieves the best results.

Neighborhood Matching Network for Entity Alignment [Website][PDF]
Yuting Wu, Xiao Liu, Yansong Feng, Zheng Wang, and Dongyan Zhao 8:00–9:00

Structural heterogeneity between knowledge graphs is an outstanding challenge for entity alignment. This paper presents Neighborhood Matching Network (NMN), a novel entity alignment framework for tackling the structural heterogeneity challenge. NMN estimates the similarities between entities to capture both the topological structure and the neighborhood difference. It provides two innovative components for better learning representations for entity alignment. It first uses a novel graph sampling method to distill a discriminative neighborhood for each entity. It then adopts a cross-graph neighborhood matching module to jointly encode the neighborhood difference for a given entity pair. Such strategies allow NMN to effectively construct matching-oriented entity representations while ignoring noisy neighbors that have a negative impact on the alignment task. Extensive experiments performed on three entity alignment datasets show that NMN can well estimate the neighborhood similarity in more tough cases and significantly outperforms 12 previous state-of-the-art methods.

Pyramid: A Layered Model for Nested Named Entity Recognition [Website][PDF]
Jue WANG, Lidan Shou, Ke Chen, and Gang Chen 8:00–9:00

This paper presents Pyramid, a novel layered model for Nested Named Entity Recognition (nested NER). In our approach, token or text region embeddings are recursively inputted into L flat NER layers, from bottom to top, stacked in a pyramid shape. Each time an embedding passes through a layer of the pyramid, its length is reduced by one. Its hidden state at layer l represents an l-gram in the input text, which is labeled only if its corresponding text region represents a complete entity mention. We also design an inverse pyramid to allow bidirectional interaction between layers. The proposed method achieves state-of-the-art F1 scores in nested NER on ACE-2004, ACE-2005, GENIA, and NNE, which are 80.27, 79.42, 77.78, and 93.70 with conventional embeddings, and 87.74, 86.34, 79.31, and 94.68 with pre-trained contextualized embeddings. In addition, our model can be used for the more general task of Overlapping Named Entity Recognition. A preliminary experiment confirms the effectiveness of our method in overlapping NER.

ReInceptionE: Relation-Aware Inception Network with Joint Local-Global Structural Information for Knowledge Graph Embedding [Website][PDF]
Zhiwen Xie, Guangyou Zhou, Jin Liu, and Jimmy Xiangji Huang 8:00–9:00

The goal of Knowledge graph embedding (KGE) is to learn how to represent the low dimensional vectors for entities and relations based on the observed triples. The conventional shallow models are limited to their expressiveness. ConvE (Dettmers et al., 2018) takes advantage of CNN and improves the expressive power with parameter efficient operators by increasing the interactions between head and relation embeddings. However, there is no structural information in the embedding space of ConvE, and the performance is still limited by the number of interactions. The recent KBGAT (Nathani et al., 2019) provides another way to learn embeddings by adaptively utilizing structural information. In this paper, we take the benefits of ConvE and KBGAT together and propose a Relation-aware Inception network with joint local-global structural information for knowledge graph Embedding (ReInceptionE). Specifically, we first explore the Inception network to learn query embedding, which aims to further increase the interactions between head and relation embeddings. Then, we propose to use a relation-aware attention mechanism to enrich the query embedding with the local neighborhood and global entity information. Experimental results on both WN18RR and FB15k-237 datasets demonstrate that ReInceptionE achieves competitive performance compared with state-of-the-art methods.

Relabel the Noise: Joint Extraction of Entities and Relations via Cooperative Multiagents [Website][PDF]

Daoyuan Chen, Yaliang Li, Kai Lei, and Ying Shen 8:00–9:00

Distant supervision based methods for entity and relation extraction have received increasing popularity due to the fact that these methods require light human annotation efforts. In this paper, we consider the problem of shifted label distribution, which is caused by the inconsistency between the noisy-labeled training set subject to external knowledge graph and the human-annotated test set, and exacerbated by the pipelined entity-then-relation extraction manner with noise propagation. We propose a joint extraction approach to address this problem by re-labeling noisy instances with a group of cooperative multiagents. To handle noisy instances in a fine-grained manner, each agent in the cooperative group evaluates the instance by calculating a continuous confidence score from its own perspective; To leverage the correlations between these two extraction tasks, a confidence consensus module is designed to gather the wisdom of all agents and re-distribute the noisy training set with confidence-scored labels. Further, the confidences are used to adjust the training losses of extractors. Experimental results on two real-world datasets verify the benefits of re-labeling noisy instance, and show that the proposed model significantly outperforms the state-of-the-art entity and relation extraction methods.

Simplify the Usage of Lexicon in Chinese NER [Website][PDF]

Ruotian Ma, Minlong Peng, Qi Zhang, Zhongyu Wei, and Xuanjing Huang 8:00–9:00

Recently, many works have tried to augment the performance of Chinese named entity recognition (NER) using word lexicons. As a representative, Lattice-LSTM has achieved new benchmark results on several public Chinese NER datasets. However, Lattice-LSTM has a complex model architecture. This limits its application in many industrial areas where real-time NER responses are needed. In this work, we propose a simple but effective method for incorporating the word lexicon into the character representations. This method avoids designing a complicated sequence modeling architecture, and for any neural NER model, it requires only subtle adjustment of the character representation layer to introduce the lexicon information. Experimental studies on four benchmark Chinese NER datasets show that our method achieves an inference speed up to 6.15 times faster than those of state-of-the-art methods, along with a better performance. The experimental results also show that the proposed method can be easily incorporated with pre-trained models like BERT.

Session 12A: Information Retrieval and Text Mining-6

Document Translation vs. Query Translation for Cross-Lingual Information Retrieval in the Medical Domain [Website][PDF]
Shadi Saleh and Pavel Pecina 8:00–9:00

We present a thorough comparison of two principal approaches to Cross-Lingual Information Retrieval: document translation (DT) and query translation (QT). Our experiments are conducted using the cross-lingual test collection produced within the CLEF eHealth information retrieval tasks in 2013—2015 containing English documents and queries in several European languages. We exploit the Statistical Machine Translation (SMT) and Neural Machine Translation (NMT) paradigms and train several domain-specific and task-specific machine translation systems to translate the non-English queries into English (for the QT approach) and the English documents to all the query languages (for the DT approach). The results show that the quality of QT by SMT is sufficient enough to outperform the retrieval results of the DT approach for all the languages. NMT then further boosts translation quality and retrieval quality for both QT and DT for most languages, but still, QT provides generally better retrieval results than DT.

Learning Robust Models for e-Commerce Product Search [Website][PDF]
Thanh Nguyen, Nikhil Rao, and Karthik Subbian 8:00–9:00

Showing items that do not match search query intent degrades customer experience in e-commerce. These mismatches result from counterfactual biases of the ranking algorithms toward noisy behavioral signals such as clicks and purchases in the search logs. Mitigating the problem requires a large labeled dataset, which is expensive and time-consuming to obtain. In this paper, we develop a deep, end-to-end model that learns to effectively classify mismatches and to generate hard mismatched examples to improve the classifier. We train the model end-to-end by introducing a latent variable into the cross-entropy loss that alternates between using the real and generated samples. This not only makes the classifier more robust but also boosts the overall ranking performance. Our model achieves a relative gain compared to baselines by over $26%$ in F-score, and over $17%$ in Area Under PR curve. On live search traffic, our model gains significant improvement in multiple countries.

Session 12A: Machine Learning for NLP-13

Dependency Graph Enhanced Dual-transformer Structure for Aspect-based Sentiment Classification [Website][PDF]
Hao Tang, Donghong Ji, Chenliang Li, and Qiji Zhou 8:00–9:00

Aspect-based sentiment classification is a popular task aimed at identifying the corresponding emotion of a specific aspect. One sentence may contain various sentiments for different aspects. Many sophisticated methods such as attention mechanism and Convolutional Neural Networks (CNN) have been widely employed for handling this challenge. Recently, semantic dependency tree implemented by Graph Convolutional Networks (GCN) is introduced to describe the inner connection between aspects and the associated emotion words. But the improvement is limited due to the noise and instability of dependency trees. To this end, we propose a dependency graph enhanced dual-transformer network (named DGEDT) by jointly considering the flat representations learnt from Transformer and graph-based representations learnt from the corresponding dependency graph in an iterative interaction manner. Specifically, a dual-transformer structure is devised in DGEDT to support mutual reinforcement between the flat representation learning and graph-based representation learning. The idea is to allow the dependency graph to guide the representation learning of the transformer encoder and vice versa. The results on five datasets demonstrate that the proposed DGEDT outperforms all state-of-the-art alternatives with a large margin.

Differentiable Window for Dynamic Local Attention [Website][PDF]
Thanh-Tung Nguyen, Xuan-Phi Nguyen, Shafiq Joty, and Xiaoli Li 8:00–9:00

We propose Differentiable Window, a new neural module and general purpose component for dynamic window selection. While universally applicable, we demonstrate a compelling use case of utilizing Differentiable Window to improve standard attention modules by enabling more focused attentions over the input regions. We propose two variants of Differentiable Window, and integrate them within the Transformer architecture in two novel ways. We evaluate our proposed approach on a myriad of NLP tasks, including machine translation, sentiment analysis, subject-verb agreement and language modeling. Our experimental results demonstrate consistent and sizable improvements across all tasks.

Generalized Entropy Regularization or: There's Nothing Special about Label Smoothing [Website][PDF]
Clara Meister, Elizabeth Salesky, and Ryan Cotterell 8:00–9:00

Prior work has explored directly regularizing the output distributions of probabilistic models to alleviate peaky (i.e. over-confident) predictions, a common sign of overfitting. This class of techniques, of which label smoothing is one, has a connection to entropy regularization. Despite the consistent success of label smoothing across architectures and data sets in language generation tasks, two problems remain open: (1) there is little understanding of the underlying effects entropy regularizers have on models, and (2) the full space of entropy regularization techniques is largely unexplored. We introduce a parametric family of entropy regularizers, which includes label smoothing as a special case, and use it to gain a better understanding of the relationship between the entropy of a model and its performance on language generation tasks. We also find that variance in model performance can be explained largely by the resulting entropy of the model. Lastly, we find that label smoothing provably does not allow for sparsity in an output distribution, an undesirable property for language generation models, and therefore advise the use of other entropy regularization methods in its place.

Highway Transformer: Self-Gating Enhanced Self-Attentive Networks [Website][PDF]
Yekun Chai, Shuo Jin, and Xinwen Hou 8:00–9:00

Self-attention mechanisms have made striking state-of-the-art (SOTA) progress in various sequence learning tasks, standing on the multi-headed dot product attention by attending to all the global contexts at different locations. Through a pseudo information highway, we introduce a gated component self-dependency units (SDU) that incorporates LSTM-styled gating units to replenish internal semantic importance within the multi-dimensional latent space of individual representations. The subsidiary content-based SDU gates allow for the information flow of modulated latent embeddings through skipped connections, leading to a clear margin of convergence speed with gradient descent algorithms. We may unveil the role of gating mechanism to aid in the context-based Transformer modules, with hypothesizing that SDU gates, especially on shallow layers, could push it faster to step towards suboptimal points during the optimization process.

Low-Dimensional Hyperbolic Knowledge Graph Embeddings [Website][PDF]
Ines Chami, Adva Wolf, Da-Cheng Juan, Frederic Sala, Sujith Ravi, and Christopher Ré 8:00–9:00

Knowledge graph (KG) embeddings learn low- dimensional representations of entities and relations to predict missing facts. KGs often exhibit hierarchical and logical patterns which must be preserved in the embedding space. For hierarchical data, hyperbolic embedding methods have shown promise for high-fidelity and parsimonious representations. However, existing hyperbolic embedding methods do not account for the rich logical patterns in KGs. In this work, we introduce a class of hyperbolic KG embedding models that simultaneously capture hierarchical and logical patterns. Our approach combines hyperbolic reflections and rotations with attention to model complex relational patterns. Experimental results on standard KG benchmarks show that our method improves over previous Euclidean- and hyperbolic-based efforts by up to 6.1% in mean reciprocal rank (MRR) in low dimensions. Furthermore, we observe that different geometric transformations capture different types of relations while attention- based transformations generalize to multiple relations. In high dimensions, our approach yields new state-of-the-art MRRs of 49.6% on WN18RR and 57.7% on YAGO3-10.

Session 12A: Machine Translation-14

AdvAug: Robust Adversarial Augmentation for Neural Machine Translation [Website][PDF]
Yong Cheng, Lu Jiang, Wolfgang Macherey, and Jacob Eisenstein 8:00–9:00

In this paper, we propose a new adversarial augmentation method for Neural Machine Translation (NMT). The main idea is to minimize the vicinal risk over virtual sentences sampled from two vicinity distributions, in which the crucial one is a novel vicinity distribution for adversarial sentences that describes a smooth interpolated embedding space centered around observed training sentence pairs. We then discuss our approach, AdvAug, to train NMT models using the embeddings of virtual sentences in sequence-to-sequence learning. Experiments on Chinese-English, English-French, and English-German translation benchmarks show that AdvAug achieves significant improvements over theTransformer (up to 4.9 BLEU points), and substantially outperforms other data augmentation techniques (e.g.back-translation) without using extra corpora.

Classification-Based Self-Learning for Weakly Supervised Bilingual Lexicon Induction [Website][PDF]
Mladen Karan, Ivan Vulić, Anna Korhonen, and Goran Glavaš 8:00–9:00

Effective projection-based cross-lingual word embedding (CLWE) induction critically relies on the iterative self-learning procedure. It gradually expands the initial small seed dictionary to learn improved cross-lingual mappings. In this work, we present ClassyMap, a classification-based approach to self-learning, yielding a more robust and a more effective induction of projection-based CLWEs. Unlike prior self-learning methods, our approach allows for integration of diverse features into the iterative process. We show the benefits of ClassyMap for bilingual lexicon induction: we report consistent improvements in a weakly supervised setup (500 seed translation pairs) on a benchmark with 28 language pairs.

Contextual Neural Machine Translation Improves Translation of Cataphoric Pronouns [Website][PDF]
KayYen Wong, Sameen Maruf, and Gholamreza Haffari 8:00–9:00

The advent of context-aware NMT has resulted in promising improvements in the overall translation quality and specifically in the translation of discourse phenomena such as pronouns. Previous works have mainly focused on the use of past sentences as context with a focus on anaphora translation. In this work, we investigate the effect of future sentences as context by comparing the performance of a contextual NMT model trained with the future context to the one trained with the past context. Our experiments and evaluation, using generic and pronoun-focused automatic metrics, show that the use of future context not only achieves significant improvements over the context-agnostic Transformer, but also demonstrates comparable and in some cases improved performance over its counterpart trained on past context. We also perform an evaluation on a targeted cataphora test suite and report significant gains over the context-agnostic Transformer in terms of BLEU.

Gender in Danger? Evaluating Speech Translation Technology on the MuST-SHE Corpus [Website][PDF]
Luisa Bentivogli, Beatrice Savoldi, Matteo Negri, Mattia A. Di Gangi, Roldano Cattoni, and Marco Turchi 8:00–9:00

Translating from languages without productive grammatical gender like English into gender-marked languages is a well-known difficulty for machines. This difficulty is also due to the fact that the training data on which models are built typically reflect the asymmetries of natural languages, gender bias included. Exclusively fed with textual data, machine translation is intrinsically constrained by the fact that the input sentence does not always contain clues about the gender identity of the referred human entities. But what happens with speech translation, where the input is an audio signal? Can audio provide additional information to reduce gender bias? We present the first thorough investigation of gender bias in speech translation, contributing with: i) the release of a benchmark useful for future studies, and ii) the comparison of different technologies (cascade and end-to-end) on two language directions (English-Italian/French).

Improving Neural Machine Translation with Soft Template Prediction [Website][PDF]
Jian Yang, Shuming Ma, Dongdong Zhang, Zhoujun Li, and Ming Zhou 8:00–9:00

Although neural machine translation (NMT) has achieved significant progress in recent years, most previous NMT models only depend on the source text to generate translation. Inspired by the success of template-based and syntax-based approaches in other fields, we propose to use extracted templates from tree structures as soft target templates to guide the translation procedure. In order to learn the syntactic structure of the target sentences, we adopt constituency-based parse tree to generate candidate templates. We incorporate the template information into the encoder-decoder framework to jointly utilize the templates and source text. Experiments show that our model significantly outperforms the baseline models on four benchmarks and demonstrates the effectiveness of soft target templates.

Uncertainty-Aware Curriculum Learning for Neural Machine Translation [Website][PDF]
Yikai Zhou, Baosong Yang, Derek F. Wong, Yu Wan, and Lidia S. Chao 8:00–9:00

Neural machine translation (NMT) has proven to be facilitated by curriculum learning which presents examples in an easy-to-hard order at different training stages. The keys lie in the assessment of data difficulty and model competence. We propose uncertainty-aware curriculum learning, which is motivated by the intuition that: 1) the higher the uncertainty in a translation pair, the more complex and rarer the information it contains; and 2) the end of the decline in model uncertainty indicates the completeness of current training stage. Specifically, we serve cross-entropy of an example as its data difficulty and exploit the variance of distributions over the weights of the network to present

the model uncertainty. Extensive experiments on various translation tasks reveal that our approach outperforms the strong baseline and related methods on both translation quality and convergence speed. Quantitative analyses reveal that the proposed strategy offers NMT the ability to automatically govern its learning schedule.

Worse WER, but Better BLEU? Leveraging Word Embedding as Intermediate in Multitask End-to-End Speech Translation [Website][PDF]

Shun-Po Chuang, Tzu-Wei Sung, Alexander H. Liu, and Hung-yi Lee 8:00–9:00

Speech translation (ST) aims to learn transformations from speech in the source language to the text in the target language. Previous works show that multitask learning improves the ST performance, in which the recognition decoder generates the text of the source language, and the translation decoder obtains the final translations based on the output of the recognition decoder. Because whether the output of the recognition decoder has the correct semantics is more critical than its accuracy, we propose to improve the multitask ST model by utilizing word embedding as the intermediate.

Session 12A: NLP Applications-9

Closing the Gap: Joint De-Identification and Concept Extraction in the Clinical Domain [Website][PDF]
Lukas Lange, Heike Adel, and Jannik Strötgen 8:00–9:00

Exploiting natural language processing in the clinical domain requires de-identification, i.e., anonymization of personal information in texts. However, current research considers de-identification and downstream tasks, such as concept extraction, only in isolation and does not study the effects of de-identification on other tasks. In this paper, we close this gap by reporting concept extraction performance on automatically anonymized data and investigating joint models for de-identification and concept extraction. In particular, we propose a stacked model with restricted access to privacy sensitive information and a multitask model. We set the new state of the art on benchmark datasets in English (96.1% F1 for de-identification and 88.9% F1 for concept extraction) and Spanish (91.4% F1 for concept extraction).

CorefQA: Coreference Resolution as Query-based Span Prediction [Website][PDF]
Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li 8:00–9:00

In this paper, we present CorefQA, an accurate and extensible approach for the coreference resolution task. We formulate the problem as a span prediction task, like in question answering: A query is generated for each candidate mention using its surrounding context, and a span prediction module is employed to extract the text spans of the coreferences within the document using the generated query. This formulation comes with the following key advantages: (1) The span prediction strategy provides the flexibility of retrieving mentions left out at the mention proposal stage; (2) In the question answering framework, encoding the mention and its context explicitly in a query makes it possible to have a deep and thorough examination of cues embedded in the context of coreferent mentions; and (3) A plethora of existing question answering datasets can be used for data augmentation to improve the model's generalization capability. Experiments demonstrate significant performance boost over previous models, with 83.1 (+3.5) F1 score on the CoNLL-2012 benchmark and 87.5 (+2.5) F1 score on the GAP benchmark.

Estimating predictive uncertainty for rumour verification models [Website][PDF]
Elena Kochkina and Maria Liakata 8:00–9:00

The inability to correctly resolve rumours circulating online can have harmful real-world consequences. We present a method for incorporating model and data uncertainty estimates into natural language processing models for automatic rumour verification. We show that these estimates can be used to filter out model predictions likely to be erroneous so that these difficult instances can be prioritised by a human fact-checker. We propose two methods for uncertainty-based instance rejection, supervised and unsupervised. We also show how uncertainty estimates can be used to interpret model performance as a rumour unfolds.

From Zero to Hero: Human-In-The-Loop Entity Linking in Low Resource Domains [Website][PDF]
Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych 8:00–9:00

Entity linking (EL) is concerned with disambiguating entity mentions in a text against knowledge bases (KB). It is crucial in a considerable number of fields like humanities, technical writing and biomedical sciences to enrich texts with semantics and discover more knowledge. The use of EL in such domains requires handling noisy texts, low resource settings and domain-specific KBs. Existing approaches are mostly inappropriate for this, as they depend on training data. However, in the above scenario, there exists hardly annotated data, and it needs to be created from scratch. We therefore present a novel domain-agnostic Human-In-The-Loop annotation approach: we use recommenders that suggest potential concepts and adaptive candidate ranking, thereby speeding up the overall annotation process and making it less tedious for users. We evaluate our ranking approach in a simulation on difficult texts and show that it greatly outperforms a strong baseline in ranking accuracy. In a user study, the annotation speed improves by 35% compared to annotating without interactive support; users report that they strongly prefer our system. An open-source and ready-to-use implementation based on the text annotation platform INCEpTION (https://inception-project.github.io) is made available.

Language to Network: Conditional Parameter Adaptation with Natural Language Descriptions [Website][PDF]
Tian Jin, Zhun Liu, Shengjia Yan, Alexandre Eichenberger, and Louis-Philippe Morency 8:00–9:00

Transfer learning using ImageNet pre-trained models has been the de facto approach in a wide range of computer vision tasks. However, fine-tuning still requires task-specific training data. In this paper, we propose $\mathbf{N^3}$ (**N**eural **N**etworks from **N**atural Language) - a new paradigm of synthesizing task-specific neural networks from language descriptions and a generic pre-trained model. $\mathbf{N^3}$ leverages language descriptions to generate parameter adaptations as well as a new task-specific classification layer for a pre-trained neural network, effectively "fine-tuning" the network for a new task using only language descriptions as input. To the best of our knowledge, $\mathbf{N^3}$ is the first method to synthesize entire neural networks from natural language. Experimental results show that $\mathbf{N^3}$ can out-perform previous natural-language based zero-shot learning methods across 4 different zero-shot image classification benchmarks. We also demonstrate a simple method to help identify keywords in language descriptions leveraged by $\mathbf{N^3}$ when synthesizing model parameters.

Neural-DINF: A Neural Network based Framework for Measuring Document Influence [Website][PDF]
Jie Tan, Changlin Yang, Ying Li, Siliang Tang, Chen Huang, and Yueting Zhuang 8:00–9:00

Measuring the scholarly impact of a document without citations is an important and challenging problem. Existing approaches such as Document Influence Model (DIM) are based on dynamic topic models, which only consider the word frequency change. In this paper, we use both frequency changes and word semantic shifts to measure document influence by developing a neural network framework. Our model has three steps. Firstly, we train the word embeddings for different time periods. Subsequently, we propose an unsupervised method to align vectors for different time periods. Finally, we compute the influence value of documents. Our experimental results show that our model outperforms DIM.

Session 12A Semantics: Sentence Level-7

[TACL] AMR-To-Text Generation with Graph Transformer [Website][PDF]
Tianming Wang, Xiaojun Wan, and Hanqi Jin 8:00–9:00

Abstract meaning representation (AMR)-to-text generation is the challenging task of generating natural language texts from AMR graphs, where nodes represent concepts and edges denote relations. The current state-of-the-art methods use graph-to-sequence models; however, they still cannot significantly outperform the previous sequence-to-sequence models or statistical approaches. In this paper, we propose a novel graph-to-sequence model (Graph Transformer) to address the above-mentioned task. The model directly encodes the AMR graphs and learns the node representations. A pairwise interaction function is used for computing the semantic relations between the concepts. Moreover, attention mechanisms are employed for aggregating the information from the incoming and outgoing neighbors, which help the model to capture the semantic information effectively. Our model outperforms the state-of-the-art neural approach by 1.5 BLEU points on LDC2015E86 and 4.8 BLEU points on LDC2017T10 and achieves new state-of-the-art performances.

Controlled Crowdsourcing for High-Quality QA-SRL Annotation [Website][PDF]
Paul Roit, Ayal Klein, Daniela Stepanov, Jonathan Mamou, Julian Michael, Gabriel Stanovsky, Luke Zettlemoyer, and Ido Dagan 8:00–9:00

Question-answer driven Semantic Role Labeling (QA-SRL) was proposed as an attractive open and natural flavour of SRL, potentially attainable from laymen. Recently, a large-scale crowdsourced QA-SRL corpus and a trained parser were released. Trying to replicate the QA-SRL annotation for new texts, we found that the resulting annotations were lacking in quality, particularly in coverage, making them insufficient for further research and evaluation. In this paper, we present an improved crowdsourcing protocol for complex semantic annotation, involving worker selection and training, and a data consolidation phase. Applying this protocol to QA-SRL yielded high-quality annotation with drastically higher coverage, producing a new gold evaluation dataset. We believe that our annotation protocol and gold standard will facilitate future replicable research of natural semantic annotations.

Cross-Lingual Semantic Role Labeling with High-Quality Translated Training Corpus [Website][PDF]
Hao Fei, Meishan Zhang, and Donghong Ji 8:00–9:00

Many efforts of research are devoted to semantic role labeling (SRL) which is crucial for natural language understanding. Supervised approaches have achieved impressing performances when large-scale corpora are available for resource-rich languages such as English. While for the low-resource languages with no annotated SRL dataset, it is still challenging to obtain competitive performances. Cross-lingual SRL is one promising way to address the problem, which has achieved great advances with the help of model transferring and annotation projection. In this paper, we propose a novel alternative based on corpus translation, constructing high-quality training datasets for the target languages from the source gold-standard SRL annotations. Experimental results on Universal Proposition Bank show that the translation-based method is highly effective, and the automatic pseudo datasets can improve the target-language SRL performances significantly.

Semantic Parsing for English as a Second Language [Website][PDF]
Yuanyuan Zhao, Weiwei Sun, junjie cao junjie, and Xiaojun Wan 8:00–9:00

This paper is concerned with semantic parsing for English as a second language (ESL). Motivated by the theoretical emphasis on the learning challenges that occur at the syntax-semantics interface during second language acquisition, we formulate the task based on the divergence between literal and intended meanings. We combine the complementary strengths of English Resource Grammar, a linguistically-precise hand-crafted deep grammar, and TLE, an existing manually annotated ESL UD-TreeBank with a novel reranking model. Experiments demonstrate that in comparison to human annotations, our method can obtain a very promising SemBanking quality. By means of the newly created corpus, we evaluate state-of-the-art semantic parsing as well as grammatical error correction models. The evaluation profiles the performance of neural NLP techniques for handling ESL data and suggests some research directions.

Sentence Meta-Embeddings for Unsupervised Semantic Textual Similarity [Website][PDF]
Nina Poerner, Ulli Waltinger, and Hinrich Schütze 8:00–9:00

We address the task of unsupervised Semantic Textual Similarity (STS) by ensembling diverse pre-trained sentence encoders into sentence meta-embeddings. We apply, extend and evaluate different meta-embedding methods from the word embedding literature at the sentence level, including dimensionality reduction (Yin and Schütze, 2016), generalized Canonical Correlation Analysis (Rastogi et al., 2015) and cross-view auto-encoders (Bollegala and Bao, 2018). Our sentence meta-embeddings set a new unsupervised State of The Art (SoTA) on the STS Benchmark and on the STS12-STS16 datasets, with gains of between 3.7% and 6.4% Pearson's r over single-source systems.

Transition-based Semantic Dependency Parsing with Pointer Networks [Website][PDF]
Daniel Fernández-González and Carlos Gómez-Rodríguez 8:00–9:00

Transition-based parsers implemented with Pointer Networks have become the new state of the art in dependency parsing, excelling in producing labelled syntactic trees and outperforming graph-based models in this task. In order to further test the capabilities of these powerful neural networks on a harder NLP problem, we propose a transition system that, thanks to Pointer Networks, can straightforwardly produce labelled directed acyclic graphs and perform semantic dependency parsing. In addition, we enhance our approach with deep contextualized word embeddings extracted from BERT. The resulting system not only outperforms all existing transition-based models, but also matches the best fully-supervised accuracy to date on the SemEval 2015 Task 18 datasets among previous state-

of-the-art graph-based parsers.

Unsupervised Dual Paraphrasing for Two-stage Semantic Parsing [Website] [PDF]
Ruisheng Cao, Su Zhu, Chenyu Yang, Chen Liu, Rao Ma, Yanbin Zhao, Lu Chen, and Kai Yu 8:00–9:00

One daunting problem for semantic parsing is the scarcity of annotation. Aiming to reduce nontrivial human labor, we propose a two-stage semantic parsing framework, where the first stage utilizes an unsupervised paraphrase model to convert an unlabeled natural language utterance into the canonical utterance. The downstream naive semantic parser accepts the intermediate output and returns the target logical form. Furthermore, the entire training process is split into two phases: pre-training and cycle learning. Three tailored self-supervised tasks are introduced throughout training to activate the unsupervised paraphrase model. Experimental results on benchmarks Overnight and GeoGranno demonstrate that our framework is effective and compatible with supervised training.

Word-level Textual Adversarial Attacking as Combinatorial Optimization [Website] [PDF]
Yuan Zang, Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Meng Zhang, Qun Liu, and Maosong Sun 8:00–9:00

Adversarial attacks are carried out to reveal the vulnerability of deep neural networks. Textual adversarial attacking is challenging because text is discrete and a small perturbation can bring significant change to the original input. Word-level attacking, which can be regarded as a combinatorial optimization problem, is a well-studied class of textual attack methods. However, existing word-level attack models are far from perfect, largely because unsuitable search space reduction methods and inefficient optimization algorithms are employed. In this paper, we propose a novel attack model, which incorporates the sememe-based word substitution method and particle swarm optimization-based search algorithm to solve the two problems separately. We conduct exhaustive experiments to evaluate our attack model by attacking BiLSTM and BERT on three benchmark datasets. Experimental results demonstrate that our model consistently achieves much higher attack success rates and crafts more high-quality adversarial examples as compared to baseline methods. Also, further experiments show our model has higher transferability and can bring more robustness enhancement to victim models by adversarial training. All the code and data of this paper can be obtained on https://github.com/thunlp/SememePSO-Attack.

tBERT: Topic Models and BERT Joining Forces for Semantic Similarity Detection [Website] [PDF]
Nicole Peinelt, Dong Nguyen, and Maria Liakata 8:00–9:00

Semantic similarity detection is a fundamental task in natural language understanding. Adding topic information has been useful for previous feature-engineered semantic similarity models as well as neural models for other tasks. There is currently no standard way of combining topics with pretrained contextual representations such as BERT. We propose a novel topic-informed BERT-based architecture for pairwise semantic similarity detection and show that our model improves performance over strong neural baselines across a variety of English language datasets. We find that the addition of topics to BERT helps particularly with resolving domain-specific cases.

Session 12A: Sentiment Analysis, Stylistic Analysis, and Argument Mining-10

Conditional Augmentation for Aspect Term Extraction via Masked Sequence-to-Sequence Generation [Website][PDF]
Kun Li, Chengbo Chen, Xiaojun Quan, Qing Ling, and Yan Song 8:00–9:00

Aspect term extraction aims to extract aspect terms from review texts as opinion targets for sentiment analysis. One of the big challenges with this task is the lack of sufficient annotated data. While data augmentation is potentially an effective technique to address the above issue, it is uncontrollable as it may change aspect words and aspect labels unexpectedly. In this paper, we formulate the data augmentation as a conditional generation task: generating a new sentence while preserving the original opinion targets and labels. We propose a masked sequence-to-sequence method for conditional augmentation of aspect term extraction. Unlike existing augmentation approaches, ours is controllable and allows to generate more diversified sentences. Experimental results confirm that our method alleviates the data scarcity problem significantly. It also effectively boosts the performances of several current models for aspect term extraction.

Exploiting Personal Characteristics of Debaters for Predicting Persuasiveness [Website][PDF]
Khalid Al Khatib, Michael Völske, Shahbaz Syed, Nikolay Kolyada, and Benno Stein 8:00–9:00

Predicting the persuasiveness of arguments has applications as diverse as writing assistance, essay scoring, and advertising. While clearly relevant to the task, the personal characteristics of an argument's source and audience have not yet been fully exploited toward automated persuasiveness prediction. In this paper, we model debaters' prior beliefs, interests, and personality traits based on their previous activity, without dependence on explicit user profiles or questionnaires. Using a dataset of over 60,000 argumentative discussions, comprising more than three million individual posts collected from the subreddit r/ChangeMyView, we demonstrate that our modeling of debater's characteristics enhances the prediction of argument persuasiveness as well as of debaters' resistance to persuasion.

Out of the Echo Chamber: Detecting Countering Debate Speeches [Website][PDF]
Matan Orbach, Yonatan Bilu, Assaf Toledo, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim 8:00–9:00

An educated and informed consumption of media content has become a challenge in modern times. With the shift from traditional news outlets to social media and similar venues, a major concern is that readers are becoming encapsulated in "echo chambers" and may fall prey to fake news and disinformation, lacking easy access to dissenting views. We suggest a novel task aiming to alleviate some of these concerns – that of detecting articles that most effectively counter the arguments – and not just the stance – made in a given text. We study this problem in the context of debate speeches. Given such a speech, we aim to identify, from among a set of speeches on the same topic and with an opposing stance, the ones that directly counter it. We provide a large dataset of 3,685 such speeches (in English), annotated for this relation, which hopefully would be of general interest to the NLP community. We explore several algorithms addressing this task, and while some are successful, all fall short of expert human performance, suggesting room for further research. All data collected during this work is freely available for research.

Session 12A: Student Research Workshop

Zero-shot North Korean to English Neural Machine Translation by Character Tokenization and Phoneme Decomposition [Website] [PDF]
Hwichan Kim, Tosho Hirasawa, and Mamoru Komachi 8:00–9:00

The primary limitation of North Korean to English translation is the lack of a parallel corpus; therefore, high translation accuracy cannot be achieved. To address this problem, we propose a zero-shot approach using South Korean data, which are remarkably similar to North Korean data. We train a neural machine translation model after tokenizing a South Korean text at the character level and decomposing characters into phonemes.We demonstrate that our method can effectively learn North Korean to English translation and improve the BLEU scores by +1.01 points in comparison with the baseline.

Research on Task Discovery for Transfer Learning in Deep Neural Networks [Website] [PDF]
Arda Akdemir 8:00–9:00

Deep neural network based machine learning models are shown to perform poorly on unseen or out-of-domain examples by numerous recent studies. Transfer learning aims to avoid overfitting and to improve generalizability by leveraging the information obtained from multiple tasks. Yet, the benefits of transfer learning depend largely on task selection and finding the right method of sharing. In this thesis, we hypothesize that current deep neural network based transfer learning models do not achieve their fullest potential for various tasks and there are still many task combinations that will benefit from transfer learning that are not considered by the current models. To this end, we started our research by implementing a novel multi-task learner with relaxed annotated data requirements and obtained a performance improvement on two NLP tasks. We will further devise models to tackle tasks from multiple areas of machine learning, such as Bioinformatics and Computer Vision, in addition to NLP.

uBLEU: Uncertainty-Aware Automatic Evaluation Method for Open-Domain Dialogue Systems [Website] [PDF]
Tsuta Yuma, Naoki Yoshinaga, and Masashi Toyoda 8:00–9:00

Because open-domain dialogues allow diverse responses, basic reference-based metrics such as BLEU do not work well unless we prepare a massive reference set of high-quality responses for input utterances. To reduce this burden, a human-aided, uncertainty-aware metric, ΔBLEU, has been proposed; it embeds human judgment on the quality of reference outputs into the computation of multiple-reference BLEU. In this study, we instead propose a fully automatic, uncertainty-aware evaluation method for open-domain dialogue systems, *υ*BLEU. This method first collects diverse reference responses from massive dialogue data and then annotates their quality judgments by using a neural network trained on automatically collected training data. Experimental results on massive Twitter data confirmed that *υ*BLEU is comparable to ΔBLEU in terms of its correlation with human judgment and that the state of the art automatic evaluation method, RUBER, is improved by integrating *υ*BLEU.

Session 12A: Summarization-7

Automatic Generation of Citation Texts in Scholarly Papers: A Pilot Study [Website][PDF]
Xinyu Xing, Xiaosheng Fan, and Xiaojun Wan 8:00–9:00

In this paper, we study the challenging problem of automatic generation of citation texts in scholarly papers. Given the context of a citing paper A and a cited paper B, the task aims to generate a short text to describe B in the given context of A. One big challenge for addressing this task is the lack of training data. Usually, explicit citation texts are easy to extract, but it is not easy to extract implicit citation texts from scholarly papers. We thus first train an implicit citation extraction model based on BERT and leverage the model to construct a large training dataset for the citation text generation task. Then we propose and train a multi-source pointer-generator network with cross attention mechanism for citation text generation. Empirical evaluation results on a manually labeled test dataset verify the efficacy of our model. This pilot study confirms the feasibility of automatically generating citation texts in scholarly papers and the technique has the great potential to help researchers prepare their scientific papers.

Composing Elementary Discourse Units in Abstractive Summarization [Website][PDF]
Zhenwen Li, Wenhao Wu, and Sujian Li 8:00–9:00

In this paper, we argue that elementary discourse unit (EDU) is a more appropriate textual unit of content selection than the sentence unit in abstractive summarization. To well handle the problem of composing EDUs into an informative and fluent summary, we propose a novel summarization method that first designs an EDU selection model to extract and group informative EDUs and then an EDU fusion model to fuse the EDUs in each group into one sentence. We also design the reinforcement learning mechanism to use EDU fusion results to reward the EDU selection action, boosting the final summarization performance. Experiments on CNN/Daily Mail have demonstrated the effectiveness of our model.

Extractive Summarization as Text Matching [Website][PDF]
Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang 8:00–9:00

This paper creates a paradigm shift with regard to the way we build neural extractive summarization systems. Instead of following the commonly used framework of extracting sentences individually and modeling the relationship between sentences, we formulate the extractive summarization task as a semantic text matching problem, in which a source document and candidate summaries will be (extracted from the original text) matched in a semantic space. Notably, this paradigm shift to semantic matching framework is well-grounded in our comprehensive analysis of the inherent gap between sentence-level and summary-level extractors based on the property of the dataset. Besides, even instantiating the framework with a simple form of a matching model, we have driven the state-of-the-art extractive result on CNN/DailyMail to a new level (44.41 in ROUGE-1). Experiments on the other five datasets also show the effectiveness of the matching framework. We believe the power of this matching-based summarization framework has not been fully exploited. To encourage more instantiations in the future, we have released our codes, processed dataset, as well as generated summaries in `https://github.com/maszhongming/MatchSum`.

Heterogeneous Graph Neural Networks for Extractive Document Summarization [Website][PDF]
Danqing Wang, Pengfei Liu, Yining Zheng, Xipeng Qiu, and Xuanjing Huang 8:00–9:00

As a crucial step in extractive document summarization, learning cross-sentence relations has been explored by a plethora of approaches. An intuitive way is to put them in the graph-based neural network, which has a more complex structure for capturing inter-sentence relationships. In this paper, we present a heterogeneous graph-based neural network for extractive summarization (HETERSUMGRAPH), which contains semantic nodes of different granularity levels apart from sentences. These additional nodes act as the intermediary between sentences and enrich the cross-sentence relations. Besides, our graph structure is flexible in natural extension from a single-document setting to multi-document via introducing document nodes. To our knowledge, we are the first one to introduce different types of nodes into graph-based neural networks for extractive document summarization and perform a comprehensive qualitative analysis to investigate their benefits. The code will be released on Github.

Jointly Learning to Align and Summarize for Neural Cross-Lingual Summarization [Website][PDF]
Yue Cao, Hui Liu, and Xiaojun Wan 8:00–9:00

Cross-lingual summarization is the task of generating a summary in one language given a text in a different language. Previous works on cross-lingual summarization mainly focus on using pipeline methods or training an end-to-end model using the translated parallel data. However, it is a big challenge for the model to directly learn cross-lingual summarization as it requires learning to understand different languages and learning how to summarize at the same time. In this paper, we propose to ease the cross-lingual summarization training by jointly learning to align and summarize. We design relevant loss functions to train this framework and propose several methods to enhance the isomorphism and cross-lingual transfer between languages. Experimental results show that our model can outperform competitive models in most cases. In addition, we show that our model even has the ability to generate cross-lingual summaries without access to any cross-lingual corpus.

Leveraging Graph to Improve Abstractive Multi-Document Summarization [Website][PDF]
Wei Li, Xinyan Xiao, Jiachen Liu, Hua Wu, Haifeng Wang, and Junping Du 8:00–9:00

Graphs that capture relations between textual units have great benefits for detecting salient information from multiple documents and generating overall coherent summaries. In this paper, we develop a neural abstractive multi-document summarization (MDS) model which can leverage well-known graph representations of documents such as similarity graph and discourse graph, to more effectively process multiple input documents and produce abstractive summaries. Our model utilizes graphs to encode documents in order to capture cross-document relations, which is

crucial to summarizing long documents. Our model can also take advantage of graphs to guide the summary generation process, which is beneficial for generating coherent and concise summaries. Furthermore, pre-trained language models can be easily combined with our model, which further improve the summarization performance significantly. Empirical results on the WikiSum and MultiNews dataset show that the proposed architecture brings substantial improvements over several strong baselines.

Multi-Granularity Interaction Network for Extractive and Abstractive Multi-Document Summarization [Website] [PDF]

Hanqi Jin, Tianming Wang, and Xiaojun Wan 8:00–9:00

In this paper, we propose a multi-granularity interaction network for extractive and abstractive multi-document summarization, which jointly learn semantic representations for words, sentences, and documents. The word representations are used to generate an abstractive summary while the sentence representations are used to produce an extractive summary. We employ attention mechanisms to interact between different granularity of semantic representations, which helps to capture multi-granularity key information and improves the performance of both abstractive and extractive summarization. Experiment results show that our proposed model substantially outperforms all strong baseline methods and achieves the best results on the Multi-News dataset.

Demo Session 2B

Time: *8:45–9:30*

MMPE: A Multi-Modal Interface using Handwriting, Touch Reordering, and Speech Commands for Post-Editing Machine Translation [Website][PDF]

Nico Herbig, Santanu Pal, Tim Düwel, Kalliopi Meladaki, Mahsa Monshizadeh, Vladislav Hnatovskiy, Antonio Krüger, and Josef van Genabith

The shift from traditional translation to post-editing (PE) of machine-translated (MT) text can save time and reduce errors, but it also affects the design of translation interfaces, as the task changes from mainly generating text to correcting errors within otherwise helpful translation proposals. Since this paradigm shift offers potential for modalities other than mouse and keyboard, we present MMPE, the first prototype to combine traditional input modes with pen, touch, and speech modalities for PE of MT. Users can directly cross out or hand-write new text, drag and drop words for reordering, or use spoken commands to update the text in place. All text manipulations are logged in an easily interpretable format to simplify subsequent translation process research. The results of an evaluation with professional translators suggest that pen and touch interaction are suitable for deletion and reordering tasks, while speech and multi-modal combinations of select & speech are considered suitable for replacements and insertions. Overall, experiment participants were enthusiastic about the new modalities and saw them as useful extensions to mouse & keyboard, but not as a complete substitute.

Session 12B Overview – Wednesday, July 8, 2020 9:00–10:00

Track A *Dialogue and Interactive Systems-15* Abstracts	Data Manipulation: Towards Effective Instance Learning for Neural Dialogue Generation via Learning to Augment and Reweight *Cai, Chen, Song, Zhang, Zhao, and Yin* [Website][PDF]	Diverse and Informative Dialogue Generation with Context-Specific Commonsense Knowledge Awareness *Wu, Li, Zhang, Zhou, and Wu* [Website][PDF]	Diversifying Dialogue Generation with Non-Conversational Text *Su, Shen, Zhao, Xiao, Hu, Niu, and Zhou* [Website][PDF]	Generate, Delete and Rewrite: A Three-Stage Framework for Improving Persona Consistency of Dialogue Generation *Song, Wang, Zhang, Liu, and Liu* [Website][PDF]	KdConv: A Chinese Multi-domain Dialogue Dataset Towards Multi-turn Knowledge-driven Conversation *Zhou, Zheng, Huang, Huang, and Zhu* [Website][PDF]
	Learning Efficient Dialogue Policy from Demonstrations through Shaping *Wang, Peng, and Wong* [Website][PDF]	Meta-Reinforced Multi-Domain State Generator for Dialogue Systems *Huang, Feng, Hu, Wu, Du, and Ma* [Website][PDF]	Modeling Long Context for Task-Oriented Dialogue State Generation *Quan and Xiong* [Website][PDF]	Multi-Domain Dialogue Acts and Response Co-Generation *Wang, Tian, Wang, Quan, and Yu* [Website][PDF]	SAS: Dialogue State Tracking via Slot Attention and Slot Information Sharing *Hu, Yang, Chen, and Yu* [Website][PDF]
	Video-Grounded Dialogues with Pretrained Generation Language Models *Le and Hoi* [Website][PDF]				
Track B *Generation-11* Abstracts	Exploring Contextual Word-level Style Relevance for Unsupervised Style Transfer *Zhou, Chen, Liu, Xiao, Su, Guo, and Wu* [Website][PDF]	Heterogeneous Graph Transformer for Graph-to-Sequence Learning *Yao, Wang, and Wan* [Website][PDF]	Neural Data-to-Text Generation via Jointly Learning the Segmentation and Correspondence *Shen, Chang, Su, Niu, and Klakow* [Website][PDF]		
Track C *Information Extraction-6* Abstracts	An Effective Transition-based Model for Discontinuous NER *Dai, Karimi, Hachey, and Paris* [Website][PDF]	Connecting Embeddings for Knowledge Graph Entity Typing *Zhao, Xie, Liu, and WANG* [Website][PDF]	Handling Rare Entities for Neural Sequence Labeling *Li, Li, Yao, and Li* [Website][PDF]	Instance-Based Learning of Span Representations: A Case Study through Named Entity Recognition *Ouchi, Suzuki, Kobayashi, Yokoi, Kuribayashi, Konno, and Inui* [Website][PDF]	Named Entity Recognition as Dependency Parsing *Yu, Bohnet, and Poesio* [Website][PDF]
Track D *Language Grounding to Vision, Robotics and Beyond-6* Abstracts	Aligned Dual Channel Graph Convolutional Network for Visual Question Answering *Huang, Wei, Cai, Zheng, Chen, Leung, and Li* [Website][PDF]	Cross-modal Coherence Modeling for Caption Generation *Alikhani, Sharma, Li, Soricut, and Stone* [Website][PDF]	Multimodal Neural Graph Memory Networks for Visual Question Answering *Khademi* [Website][PDF]	Refer360°: A Referring Expression Recognition Dataset in 360° Images *Cirik, Berg-Kirkpatrick, and Morency* [Website][PDF]	Span-based Localizing Network for Natural Language Video Localization *Zhang, Sun, Jing, and Zhou* [Website][PDF]

Track E *Machine Learning for NLP-14* Abstracts	CamemBERT: a Tasty French Language Model *Martin, Muller, Ortiz Suárez, Dupont, Romary, Clergerie, Seddah, and Sagot* [Website] [PDF]	Effective Estimation of Deep Generative Language Models *Pelsmaeker and Aziz* [Website] [PDF]	Evaluating and Enhancing the Robustness of Neural Network-based Dependency Parsing Models with Adversarial Examples *Zheng, Zeng, Zhou, Hsieh, Cheng, and Huang* [Website] [PDF]	Learning Architectures from an Extended Search Space for Language Modeling *Li, Hu, Zhang, Xu, Jiang, Xiao, Zhu, Liu, and* [Website] [PDF]	Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection *Ravfogel, Elazar, Gonen, Twiton, and Goldberg* [Website] [PDF]
Track F *Phonology, Morphology and Word Segmentation-4* Abstracts	2kenize: Tying Subword Sequences for Chinese Script Conversion *Pranav A and Augenstein* [Website] [PDF]	Bootstrapping Techniques for Polysynthetic Morphological Analysis *Lane and Bird* [Website] [PDF]	Coupling Distant Annotation and Adversarial Training for Cross-Domain Chinese Word Segmentation *Ding, Long, Xu, Zhu, Xie, Wang, and Zheng* [Website] [PDF]	Modeling Morphological Typology for Unsupervised Learning of Language Morphology *Xu, Kodner, Marcus, and Yang* [Website] [PDF]	Predicting Declension Class from Form and Meaning *Williams, Pimentel, Blix, McCarthy, Chodroff, and Cotterell* [Website] [PDF]
	Predicting the Growth of Morphological Families from Social and Linguistic Factors *Hofmann, Pierrehumbert, and Schütze* [Website] [PDF]	Semi-supervised Contextual Historical Text Normalization *Makarov and Clematide* [Website] [PDF]			
Track G *Question Answering-10* Abstracts	ClarQ: A large-scale and diverse dataset for Clarification Question Generation *Kumar and Black* [Website] [PDF]	DoQA - Accessing Domain-Specific FAQs via Conversational QA *Campos, Otegi, Soroa, Deriu, Cieliebak, and Agirre* [Website] [PDF]	Harvesting and Refining Question-Answer Pairs for Unsupervised QA *Li, Wang, Dong, Wei, and Xu* [Website] [PDF]	MLQA: Evaluating Cross-lingual Extractive Question Answering *Lewis, Oguz, Rinott, Riedel, and Schwenk* [Website] [PDF]	Multi-source Meta Transfer for Low Resource Multiple-Choice Question Answering *Yan, Zhang, Jin, and Zhou* [Website] [PDF]
	RikiNet: Reading Wikipedia Pages for Natural Question Answering *Liu, Gong, Fu, Yan, Chen, Jiang, Lv, and Duan* [Website] [PDF]				
Track H *Textual Inference and Other Areas of Semantics-4* Abstracts	Do Neural Models Learn Systematicity of Monotonicity Inference in Natural Language? *Yanaka, Mineshima, Bekki, and Inui* [Website] [PDF]	Fine-grained Fact Verification with Kernel Graph Attention Network *Liu, Xiong, Sun, and Liu* [Website] [PDF]	Generating Fact Checking Explanations *Atanasova, Simonsen, Lioma, and Augenstein* [Website] [PDF]	How to Ask Good Questions? Try to Leverage Paraphrases *Jia, Zhou, SUN, and Wu* [Website] [PDF]	Premise Selection in Natural Language Mathematical Texts *Ferreira and Freitas* [Website] [PDF]
Track I *Student Research Workshop* Abstracts	Self-Attention is Not Only a Weight: Analyzing BERT with Vector Norms *Kobayashi, Kuribayashi, Yokoi, and Inui* [Website]	Transferring Monolingual Model to Low-Resource Language: The Case of Tigrinya *Tela, Zewoudie, and Hautamäki* [Website]	Adaptive Transformers for Learning Multimodal Representations *Bhargava* [Website] [PDF]		

Track J *Theme-4* Abstracts	A Call for More Rigor in Unsupervised Cross-lingual Learning *Artetxe, Ruder, Yogatama, Labaka, and Agirre* [Website] [PDF]	A Tale of a Probe and a Parser *Hall Maudslay, Valvoda, Pimentel, Williams, and Cotterell* [Website] [PDF]	From SPMRL to NMRL: What Did We Learn (and Unlearn) in a Decade of Parsing Morphologically-Rich Languages (MRLs)? *Tsarfaty, Bareket, Klein, and Seker* [Website] [PDF]	Speech Translation and the End-to-End Promise: Taking Stock of Where We Are *Sperber and Paulik* [Website] [PDF]	The State and Fate of Linguistic Diversity and Inclusion in the NLP World *Joshi, Santy, Budhiraja, Bali, and Choudhury* [Website] [PDF]
	What Question Answering can Learn from Trivia Nerds *Boyd-Graber and Börschinger* [Website] [PDF]	What are the Goals of Distributional Semantics? *Emerson* [Website] [PDF]			

Session 12B Details

Session 12B: Dialogue and Interactive Systems-15

Data Manipulation: Towards Effective Instance Learning for Neural Dialogue Generation via Learning to Augment and Reweight [Website][PDF]
Hengyi Cai, Hongshen Chen, Yonghao Song, Cheng Zhang, Xiaofang Zhao, and Dawei Yin 9:00–10:00

Current state-of-the-art neural dialogue models learn from human conversations following the data-driven paradigm. As such, a reliable training corpus is the crux of building a robust and well-behaved dialogue model. However, due to the open-ended nature of human conversations, the quality of user-generated training data varies greatly, and effective training samples are typically insufficient while noisy samples frequently appear. This impedes the learning of those data-driven neural dialogue models. Therefore, effective dialogue learning requires not only more reliable learning samples, but also fewer noisy samples. In this paper, we propose a data manipulation framework to proactively reshape the data distribution towards reliable samples by augmenting and highlighting effective learning samples as well as reducing the effect of inefficient samples simultaneously. In particular, the data manipulation model selectively augments the training samples and assigns an importance weight to each instance to reform the training data. Note that, the proposed data manipulation framework is fully data-driven and learnable. It not only manipulates training samples to optimize the dialogue generation model, but also learns to increase its manipulation skills through gradient descent with validation samples. Extensive experiments show that our framework can improve the dialogue generation performance with respect to various automatic evaluation metrics and human judgments.

Diverse and Informative Dialogue Generation with Context-Specific Commonsense Knowledge Awareness [Website][PDF]
Sixing Wu, Ying Li, Dawei Zhang, Yang Zhou, and Zhonghai Wu 9:00–10:00

Generative dialogue systems tend to produce generic responses, which often leads to boring conversations. For alleviating this issue, Recent studies proposed to retrieve and introduce knowledge facts from knowledge graphs. While this paradigm works to a certain extent, it usually retrieves knowledge facts only based on the entity word itself, without considering the specific dialogue context. Thus, the introduction of the context-irrelevant knowledge facts can impact the quality of generations. To this end, this paper proposes a novel commonsense knowledge-aware dialogue generation model, ConKADI. We design a Felicitous Fact mechanism to help the model focus on the knowledge facts that are highly relevant to the context; furthermore, two techniques, Context-Knowledge Fusion and Flexible Mode Fusion are proposed to facilitate the integration of the knowledge in the ConKADI. We collect and build a large-scale Chinese dataset aligned with the commonsense knowledge for dialogue generation. Extensive evaluations over both an open-released English dataset and our Chinese dataset demonstrate that our approach ConKADI outperforms the state-of-the-art approach CCM, in most experiments.

Diversifying Dialogue Generation with Non-Conversational Text [Website][PDF]
Hui Su, Xiaoyu Shen, Sanqiang Zhao, Zhou Xiao, Pengwei Hu, randy zhong randy, Cheng Niu, and Jie Zhou 9:00–10:00

Neural network-based sequence-to-sequence (seq2seq) models strongly suffer from the low-diversity problem when it comes to open-domain dialogue generation. As bland and generic utterances usually dominate the frequency distribution in our daily chitchat, avoiding them to generate more interesting responses requires complex data filtering, sampling techniques or modifying the training objective. In this paper, we propose a new perspective to diversify dialogue generation by leveraging *non-conversational* text. Compared with bilateral conversations, non-conversational text are easier to obtain, more diverse and cover a much broader range of topics. We collect a large-scale non-conversational corpus from multi sources including forum comments, idioms and book snippets. We further present a training paradigm to effectively incorporate these text via iterative back translation. The resulting model is tested on two conversational datasets from different domains and is shown to produce significantly more diverse responses without sacrificing the relevance with context.

Generate, Delete and Rewrite: A Three-Stage Framework for Improving Persona Consistency of Dialogue Generation [Website][PDF]
Haoyu Song, Yan Wang, Wei-Nan Zhang, Xiaojiang Liu, and Ting Liu 9:00–10:00

Maintaining a consistent personality in conversations is quite natural for human beings, but is still a non-trivial task for machines. The persona-based dialogue generation task is thus introduced to tackle the personality-inconsistent problem by incorporating explicit persona text into dialogue generation models. Despite the success of existing persona-based models on generating human-like responses, their one-stage decoding framework can hardly avoid the generation of inconsistent persona words. In this work, we introduce a three-stage framework that employs a generate-delete-rewrite mechanism to delete inconsistent words from a generated response prototype and further rewrite it to a personality-consistent one. We carry out evaluations by both human and automatic metrics. Experiments on the Persona-Chat dataset show that our approach achieves good performance.

KdConv: A Chinese Multi-domain Dialogue Dataset Towards Multi-turn Knowledge-driven Conversation [Website][PDF]
Hao Zhou, Chujie Zheng, Kaili Huang, Minlie Huang, and Xiaoyan Zhu 9:00–10:00

The research of knowledge-driven conversational systems is largely limited due to the lack of dialog data which con-

sists of multi-turn conversations on multiple topics and with knowledge annotations. In this paper, we propose a Chinese multi-domain knowledge-driven conversation dataset, KdConv, which grounds the topics in multi-turn conversations to knowledge graphs. Our corpus contains 4.5K conversations from three domains (film, music, and travel), and 86K utterances with an average turn number of 19.0. These conversations contain in-depth discussions on related topics and natural transition between multiple topics. To facilitate the following research on this corpus, we provide several benchmark models. Comparative results show that the models can be enhanced by introducing background knowledge, yet there is still a large space for leveraging knowledge to model multi-turn conversations for further research. Results also show that there are obvious performance differences between different domains, indicating that it is worth further explore transfer learning and domain adaptation. The corpus and benchmark models are publicly available.

Learning Efficient Dialogue Policy from Demonstrations through Shaping [Website][PDF]
Huimin Wang, Baolin Peng, and Kam-Fai Wong 9:00–10:00

Training a task-oriented dialogue agent with reinforcement learning is prohibitively expensive since it requires a large volume of interactions with users. Human demonstrations can be used to accelerate learning progress. However, how to effectively leverage demonstrations to learn dialogue policy remains less explored. In this paper, we present S^2Agent that efficiently learns dialogue policy from demonstrations through policy shaping and reward shaping. We use an imitation model to distill knowledge from demonstrations, based on which policy shaping estimates feedback on how the agent should act in policy space. Reward shaping is then incorporated to bonus state-actions similar to demonstrations explicitly in value space encouraging better exploration. The effectiveness of the proposed S^2Agentt is demonstrated in three dialogue domains and a challenging domain adaptation task with both user simulator evaluation and human evaluation.

Meta-Reinforced Multi-Domain State Generator for Dialogue Systems [Website][PDF]
Yi Huang, Junlan Feng, Min Hu, Xiaoting Wu, Xiaoyu Du, and Shuo Ma 9:00–10:00

A Dialogue State Tracker (DST) is a core component of a modular task-oriented dialogue system. Tremendous progress has been made in recent years. However, the major challenges remain. The state-of-the-art accuracy for DST is below 50% for a multi-domain dialogue task. A learnable DST for any new domain requires a large amount of labeled in-domain data and training from scratch. In this paper, we propose a Meta-Reinforced Multi-Domain State Generator (MERET). Our first contribution is to improve the DST accuracy. We enhance a neural model based DST generator with a reward manager, which is built on policy gradient reinforcement learning (RL) to fine-tune the generator. With this change, we are able to improve the joint accuracy of DST from 48.79% to 50.91% on the MultiWOZ corpus. Second, we explore to train a DST meta-learning model with a few domains as source domains and a new domain as target domain. We apply the model-agnostic meta-learning algorithm (MAML) to DST and the obtained meta-learning model is used for new domain adaptation. Our experimental results show this solution is able to outperform the traditional training approach with extremely less training data in target domain.

Modeling Long Context for Task-Oriented Dialogue State Generation [Website][PDF]
Jun Quan and Deyi Xiong 9:00–10:00

Based on the recently proposed transferable dialogue state generator (TRADE) that predicts dialogue states from utterance-concatenated dialogue context, we propose a multi-task learning model with a simple yet effective utterance tagging technique and a bidirectional language model as an auxiliary task for task-oriented dialogue state generation. By enabling the model to learn a better representation of the long dialogue context, our approaches attempt to solve the problem that the performance of the baseline significantly drops when the input dialogue context sequence is long. In our experiments, our proposed model achieves a 7.03% relative improvement over the baseline, establishing a new state-of-the-art joint goal accuracy of 52.04% on the MultiWOZ 2.0 dataset.

Multi-Domain Dialogue Acts and Response Co-Generation [Website][PDF]
Kai Wang, Junfeng Tian, Rui Wang, Xiaojun Quan, and Jianxing Yu 9:00–10:00

Generating fluent and informative responses is of critical importance for task-oriented dialogue systems. Existing pipeline approaches generally predict multiple dialogue acts first and use them to assist response generation. There are at least two shortcomings with such approaches. First, the inherent structures of multi-domain dialogue acts are neglected. Second, the semantic associations between acts and responses are not taken into account for response generation. To address these issues, we propose a neural co-generation model that generates dialogue acts and responses concurrently. Unlike those pipeline approaches, our act generation module preserves the semantic structures of multi-domain dialogue acts and our response generation module dynamically attends to different acts as needed. We train the two modules jointly using an uncertainty loss to adjust their task weights adaptively. Extensive experiments are conducted on the large-scale MultiWOZ dataset and the results show that our model achieves very favorable improvement over several state-of-the-art models in both automatic and human evaluations.

SAS: Dialogue State Tracking via Slot Attention and Slot Information Sharing [Website][PDF]
Jiaying Hu, Yan Yang, Chencai Chen, liang he liang, and Zhou Yu 9:00–10:00

Dialogue state tracker is responsible for inferring user intentions through dialogue history. Previous methods have difficulties in handling dialogues with long interaction context, due to the excessive information. We propose a Dialogue State Tracker with Slot Attention and Slot Information Sharing (SAS) to reduce redundant information's interference and improve long dialogue context tracking. Specially, we first apply a Slot Attention to learn a set of slot-specific features from the original dialogue and then integrate them using a slot information sharing module. Our model yields a significantly improved performance compared to previous state-of the-art models on the MultiWOZ dataset.

Video-Grounded Dialogues with Pretrained Generation Language Models [Website][PDF]
Hung Le and Steven C.H. Hoi 9:00–10:00

Pre-trained language models have shown remarkable success in improving various downstream NLP tasks due to their ability to capture dependencies in textual data and generate natural responses. In this paper, we leverage the power of pre-trained language models for improving video-grounded dialogue, which is very challenging and involves complex features of different dynamics: (1) Video features which can extend across both spatial and temporal dimensions; and (2) Dialogue features which involve semantic dependencies over multiple dialogue turns. We propose a framework by extending GPT-2 models to tackle these challenges by formulating video-grounded dialogue tasks as a sequence-to-sequence task, combining both visual and textual representation into a structured sequence, and fine-tuning a large pre-trained GPT-2 network. Our framework allows fine-tuning language models to capture dependencies across multiple modalities over different levels of information: spatio-temporal level in video and token-sentence level in dialogue context. We achieve promising improvement on the Audio-Visual Scene-Aware Dialogues (AVSD) benchmark from DSTC7, which supports a potential direction in this line of research.

Session 12B: Generation-11

Exploring Contextual Word-level Style Relevance for Unsupervised Style Transfer [Website][PDF]
Chulun Zhou, Liangyu Chen, Jiachen Liu, Xinyan Xiao, Jinsong Su, Sheng Guo, and Hua Wu 9:00–10:00

Unsupervised style transfer aims to change the style of an input sentence while preserving its original content without using parallel training data. In current dominant approaches, owing to the lack of fine-grained control on the influence from the target style, they are unable to yield desirable output sentences. In this paper, we propose a novel attentional sequence-to-sequence (Seq2seq) model that dynamically exploits the relevance of each output word to the target style for unsupervised style transfer. Specifically, we first pretrain a style classifier, where the relevance of each input word to the original style can be quantified via layer-wise relevance propagation. In a denoising auto-encoding manner, we train an attentional Seq2seq model to reconstruct input sentences and repredict word-level previously-quantified style relevance simultaneously. In this way, this model is endowed with the ability to automatically predict the style relevance of each output word. Then, we equip the decoder of this model with a neural style component to exploit the predicted wordlevel style relevance for better style transfer. Particularly, we fine-tune this model using a carefully-designed objective function involving style transfer, style relevance consistency, content preservation and fluency modeling loss terms. Experimental results show that our proposed model achieves state-of-the-art performance in terms of both transfer accuracy and content preservation.

Heterogeneous Graph Transformer for Graph-to-Sequence Learning [Website][PDF]
Shaowei Yao, Tianming Wang, and Xiaojun Wan 9:00–10:00

The graph-to-sequence (Graph2Seq) learning aims to transduce graph-structured representations to word sequences for text generation. Recent studies propose various models to encode graph structure. However, most previous works ignore the indirect relations between distance nodes, or treat indirect relations and direct relations in the same way. In this paper, we propose the Heterogeneous Graph Transformer to independently model the different relations in the individual subgraphs of the original graph, including direct relations, indirect relations and multiple possible relations between nodes. Experimental results show that our model strongly outperforms the state of the art on all four standard benchmarks of AMR-to-text generation and syntax-based neural machine translation.

Neural Data-to-Text Generation via Jointly Learning the Segmentation and Correspondence [Website][PDF]
Xiaoyu Shen, Ernie Chang, Hui Su, Cheng Niu, and Dietrich Klakow 9:00–10:00

The neural attention model has achieved great success in data-to-text generation tasks. Though usually excelling at producing fluent text, it suffers from the problem of information missing, repetition and "hallucination". Due to the black-box nature of the neural attention architecture, avoiding these problems in a systematic way is non-trivial. To address this concern, we propose to explicitly segment target text into fragment units and align them with their data correspondences. The segmentation and correspondence are jointly learned as latent variables without any human annotations. We further impose a soft statistical constraint to regularize the segmental granularity. The resulting architecture maintains the same expressive power as neural attention models, while being able to generate fully interpretable outputs with several times less computational cost. On both E2E and WebNLG benchmarks, we show the proposed model consistently outperforms its neural attention counterparts.

Session 12B: Information Extraction-6

An Effective Transition-based Model for Discontinuous NER [Website][PDF]
Xiang Dai, Sarvnaz Karimi, Ben Hachey, and Cecile Paris 9:00–10:00

Unlike widely used Named Entity Recognition (NER) data sets in generic domains, biomedical NER data sets often contain mentions consisting of discontinuous spans. Conventional sequence tagging techniques encode Markov assumptions that are efficient but preclude recovery of these mentions. We propose a simple, effective transition-based model with generic neural encoding for discontinuous NER. Through extensive experiments on three biomedical data sets, we show that our model can effectively recognize discontinuous mentions without sacrificing the accuracy on continuous mentions.

Connecting Embeddings for Knowledge Graph Entity Typing [Website][PDF]
Yu Zhao, anxiang zhang anxiang, Ruobing Xie, Kang Liu, and Xiaojie WANG 9:00–10:00

Knowledge graph (KG) entity typing aims at inferring possible missing entity type instances in KG, which is a very significant but still under-explored subtask of knowledge graph completion. In this paper, we propose a novel approach for KG entity typing which is trained by jointly utilizing local typing knowledge from existing entity type assertions and global triple knowledge in KGs. Specifically, we present two distinct knowledge-driven effective mechanisms of entity type inference. Accordingly, we build two novel embedding models to realize the mechanisms. Afterward, a joint model via connecting them is used to infer missing entity type instances, which favors inferences that agree with both entity type instances and triple knowledge in KGs. Experimental results on two real-world datasets (Freebase and YAGO) demonstrate the effectiveness of our proposed mechanisms and models for improving KG entity typing. The source code and data of this paper can be obtained from: https://github.com/Adam1679/ConnectE .

Handling Rare Entities for Neural Sequence Labeling [Website][PDF]
Yangming Li, Han Li, Kaisheng Yao, and Xiaolong Li 9:00–10:00

One great challenge in neural sequence labeling is the data sparsity problem for rare entity words and phrases. Most of test set entities appear only few times and are even unseen in training corpus, yielding large number of out-of-vocabulary (OOV) and low-frequency (LF) entities during evaluation. In this work, we propose approaches to address this problem. For OOV entities, we introduce local context reconstruction to implicitly incorporate contextual information into their representations. For LF entities, we present delexicalized entity identification to explicitly extract their frequency-agnostic and entity-type-specific representations. Extensive experiments on multiple benchmark datasets show that our model has significantly outperformed all previous methods and achieved new start-of-the-art results. Notably, our methods surpass the model fine-tuned on pre-trained language models without external resource.

Instance-Based Learning of Span Representations: A Case Study through Named Entity Recognition [Website][PDF]
Hiroki Ouchi, Jun Suzuki, Sosuke Kobayashi, Sho Yokoi, Tatsuki Kuribayashi, Ryuto Konno, and Kentaro Inui 9:00–10:00

Interpretable rationales for model predictions play a critical role in practical applications. In this study, we develop models possessing interpretable inference process for structured prediction. Specifically, we present a method of instance-based learning that learns similarities between spans. At inference time, each span is assigned a class label based on its similar spans in the training set, where it is easy to understand how much each training instance contributes to the predictions. Through empirical analysis on named entity recognition, we demonstrate that our method enables to build models that have high interpretability without sacrificing performance.

Named Entity Recognition as Dependency Parsing [Website][PDF]
Juntao Yu, Bernd Bohnet, and Massimo Poesio 9:00–10:00

Named Entity Recognition (NER) is a fundamental task in Natural Language Processing, concerned with identifying spans of text expressing references to entities. NER research is often focused on flat entities only (flat NER), ignoring the fact that entity references can be nested, as in [Bank of [China]] (Finkel and Manning, 2009). In this paper, we use ideas from graph-based dependency parsing to provide our model a global view on the input via a biaffine model (Dozat and Manning, 2017). The biaffine model scores pairs of start and end tokens in a sentence which we use to explore all spans, so that the model is able to predict named entities accurately. We show that the model works well for both nested and flat NER through evaluation on 8 corpora and achieving SoTA performance on all of them, with accuracy gains of up to 2.2 percentage points.

Session 12B: Language Grounding to Vision, Robotics and Beyond-6

Aligned Dual Channel Graph Convolutional Network for Visual Question Answering [Website][PDF]
Qingbao Huang, Jielong Wei, Yi Cai, Changmeng Zheng, Junying Chen, Ho-fung Leung, and Qing Li
9:00–10:00

Visual question answering aims to answer the natural language question about a given image. Existing graph-based methods only focus on the relations between objects in an image and neglect the importance of the syntactic dependency relations between words in a question. To simultaneously capture the relations between objects in an image and the syntactic dependency relations between words in a question, we propose a novel dual channel graph convolutional network (DC-GCN) for better combining visual and textual advantages. The DC-GCN model consists of three parts: an I-GCN module to capture the relations between objects in an image, a Q-GCN module to capture the syntactic dependency relations between words in a question, and an attention alignment module to align image representations and question representations. Experimental results show that our model achieves comparable performance with the state-of-the-art approaches.

Cross-modal Coherence Modeling for Caption Generation [Website][PDF]
Malihe Alikhani, Piyush Sharma, Shengjie Li, Radu Soricut, and Matthew Stone 9:00–10:00

We use coherence relations inspired by computational models of discourse to study the information needs and goals of image captioning. Using an annotation protocol specifically devised for capturing image–caption coherence relations, we annotate 10,000 instances from publicly-available image–caption pairs. We introduce a new task for learning inferences in imagery and text, coherence relation prediction, and show that these coherence annotations can be exploited to learn relation classifiers as an intermediary step, and also train coherence-aware, controllable image captioning models. The results show a dramatic improvement in the consistency and quality of the generated captions with respect to information needs specified via coherence relations.

Multimodal Neural Graph Memory Networks for Visual Question Answering [Website][PDF]
Mahmoud Khademi 9:00–10:00

We introduce a new neural network architecture, Multimodal Neural Graph Memory Networks (MN-GMN), for visual question answering. The MN-GMN uses graph structure with different region features as node attributes and applies a recently proposed powerful graph neural network model, Graph Network (GN), to reason about objects and their interactions in an image. The input module of the MN-GMN generates a set of visual features plus a set of encoded region-grounded captions (RGCs) for the image. The RGCs capture object attributes and their relationships. Two GNs are constructed from the input module using the visual features and encoded RGCs. Each node of the GNs iteratively computes a question-guided contextualized representation of the visual/textual information assigned to it. Then, to combine the information from both GNs, the nodes write the updated representations to an external spatial memory. The final states of the memory cells are fed into an answer module to predict an answer. Experiments show MN-GMN rivals the state-of-the-art models on Visual7W, VQA-v2.0, and CLEVR datasets.

Refer360°: A Referring Expression Recognition Dataset in 360° Images [Website][PDF]
Volkan Cirik, Taylor Berg-Kirkpatrick, and Louis-Philippe Morency 9:00–10:00

We propose a novel large-scale referring expression recognition dataset, Refer360°, consisting of 17,137 instruction sequences and ground-truth actions for completing these instructions in 360° scenes. Refer360° differs from existing related datasets in three ways. First, we propose a more realistic scenario where instructors and the followers have partial, yet dynamic, views of the scene – followers continuously modify their field-of-view (FoV) while interpreting instructions that specify a final target location. Second, instructions to find the target location consist of multiple steps for followers who will start at random FoVs. As a result, intermediate instructions are strongly grounded in object references, and followers must identify intermediate FoVs to find the final target location correctly. Third, the target locations are neither restricted to predefined objects nor chosen by annotators; instead, they are distributed randomly across scenes. This "point anywhere" approach leads to more linguistically complex instructions, as shown in our analyses. Our examination of the dataset shows that Refer360° manifests linguistically rich phenomena in a language grounding task that poses novel challenges for computational modeling of language, vision, and navigation.

Span-based Localizing Network for Natural Language Video Localization [Website][PDF]
Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou 9:00–10:00

Given an untrimmed video and a text query, natural language video localization (NLVL) is to locate a matching span from the video that semantically corresponds to the query. Existing solutions formulate NLVL either as a ranking task and apply multimodal matching architecture, or as a regression task to directly regress the target video span. In this work, we address NLVL task with a span-based QA approach by treating the input video as text passage. We propose a video span localizing network (VSLNet), on top of the standard span-based QA framework, to address NLVL. The proposed VSLNet tackles the differences between NLVL and span-based QA through a simple and yet effective query-guided highlighting (QGH) strategy. The QGH guides VSLNet to search for matching video span within a highlighted region. Through extensive experiments on three benchmark datasets, we show that the proposed VSLNet outperforms the state-of-the-art methods; and adopting span-based QA framework is a promising direction to solve NLVL.

Session 12B: Machine Learning for NLP-14

CamemBERT: a Tasty French Language Model [Website][PDF]
Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot 9:00–10:00

Pretrained language models are now ubiquitous in Natural Language Processing. Despite their success, most available models have either been trained on English data or on the concatenation of data in multiple languages. This makes practical use of such models –in all languages except English– very limited. In this paper, we investigate the feasibility of training monolingual Transformer-based language models for other languages, taking French as an example and evaluating our language models on part-of-speech tagging, dependency parsing, named entity recognition and natural language inference tasks. We show that the use of web crawled data is preferable to the use of Wikipedia data. More surprisingly, we show that a relatively small web crawled dataset (4GB) leads to results that are as good as those obtained using larger datasets (130+GB). Our best performing model CamemBERT reaches or improves the state of the art in all four downstream tasks.

Effective Estimation of Deep Generative Language Models [Website][PDF]
Tom Pelsmaeker and Wilker Aziz 9:00–10:00

Advances in variational inference enable parameterisation of probabilistic models by deep neural networks. This combines the statistical transparency of the probabilistic modelling framework with the representational power of deep learning. Yet, due to a problem known as posterior collapse, it is difficult to estimate such models in the context of language modelling effectively. We concentrate on one such model, the variational auto-encoder, which we argue is an important building block in hierarchical probabilistic models of language. This paper contributes a sober view of the problem, a survey of techniques to address it, novel techniques, and extensions to the model. To establish a ranking of techniques, we perform a systematic comparison using Bayesian optimisation and find that many techniques perform reasonably similar, given enough resources. Still, a favourite can be named based on convenience. We also make several empirical observations and recommendations of best practices that should help researchers interested in this exciting field.

Evaluating and Enhancing the Robustness of Neural Network-based Dependency Parsing Models with Adversarial Examples [Website][PDF]
Xiaoqing Zheng, Jiehang Zeng, Yi Zhou, Cho-Jui Hsieh, Minhao Cheng, and Xuanjing Huang 9:00–10:00

Despite achieving prominent performance on many important tasks, it has been reported that neural networks are vulnerable to adversarial examples. Previously studies along this line mainly focused on semantic tasks such as sentiment analysis, question answering and reading comprehension. In this study, we show that adversarial examples also exist in dependency parsing: we propose two approaches to study where and how parsers make mistakes by searching over perturbations to existing texts at sentence and phrase levels, and design algorithms to construct such examples in both of the black-box and white-box settings. Our experiments with one of state-of-the-art parsers on the English Penn Treebank (PTB) show that up to 77% of input examples admit adversarial perturbations, and we also show that the robustness of parsing models can be improved by crafting high-quality adversaries and including them in the training stage, while suffering little to no performance drop on the clean input data.

Learning Architectures from an Extended Search Space for Language Modeling [Website][PDF]
Yinqiao Li, Chi Hu, Yuhao Zhang, Nuo Xu, Yufan Jiang, Tong Xiao, Jingbo Zhu, Tongran Liu, and changliang li changliang 9:00–10:00

Neural architecture search (NAS) has advanced significantly in recent years but most NAS systems restrict search to learning architectures of a recurrent or convolutional cell. In this paper, we extend the search space of NAS. In particular, we present a general approach to learn both intra-cell and inter-cell architectures (call it ESS). For a better search result, we design a joint learning method to perform intra-cell and inter-cell NAS simultaneously. We implement our model in a differentiable architecture search system. For recurrent neural language modeling, it outperforms a strong baseline significantly on the PTB and WikiText data, with a new state-of-the-art on PTB. Moreover, the learned architectures show good transferability to other systems. E.g., they improve state-of-the-art systems on the CoNLL and WNUT named entity recognition (NER) tasks and CoNLL chunking task, indicating a promising line of research on large-scale pre-learned architectures.

Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection [Website][PDF]
Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg 9:00–10:00

The ability to control for the kinds of information encoded in neural representation has a variety of use cases, especially in light of the challenge of interpreting these models. We present Iterative Null-space Projection (INLP), a novel method for removing information from neural representations. Our method is based on repeated training of linear classifiers that predict a certain property we aim to remove, followed by projection of the representations on their null-space. By doing so, the classifiers become oblivious to that target property, making it hard to linearly separate the data according to it. While applicable for multiple uses, we evaluate our method on bias and fairness use-cases, and show that our method is able to mitigate bias in word embeddings, as well as to increase fairness in a setting of multi-class classification.

Session 12B: Phonology, Morphology and Word Segmentation-4

2kenize: Tying Subword Sequences for Chinese Script Conversion [Website][PDF]
Pranav A and Isabelle Augenstein 9:00–10:00

Simplified Chinese to Traditional Chinese character conversion is a common preprocessing step in Chinese NLP. Despite this, current approaches have insufficient performance because they do not take into account that a simplified Chinese character can correspond to multiple traditional characters. Here, we propose a model that can disambiguate between mappings and convert between the two scripts. The model is based on subword segmentation, two language models, as well as a method for mapping between subword sequences. We further construct benchmark datasets for topic classification and script conversion. Our proposed method outperforms previous Chinese Character conversion approaches by 6 points in accuracy. These results are further confirmed in a downstream application, where 2kenize is used to convert pretraining dataset for topic classification. An error analysis reveals that our method's particular strengths are in dealing with code mixing and named entities.

Bootstrapping Techniques for Polysynthetic Morphological Analysis [Website][PDF]
William Lane and Steven Bird 9:00–10:00

Polysynthetic languages have exceptionally large and sparse vocabularies, thanks to the number of morpheme slots and combinations in a word. This complexity, together with a general scarcity of written data, poses a challenge to the development of natural language technologies. To address this challenge, we offer linguistically-informed approaches for bootstrapping a neural morphological analyzer, and demonstrate its application to Kunwinjku, a polysynthetic Australian language. We generate data from a finite state transducer to train an encoder-decoder model. We improve the model by "hallucinating" missing linguistic structure into the training data, and by resampling from a Zipf distribution to simulate a more natural distribution of morphemes. The best model accounts for all instances of reduplication in the test set and achieves an accuracy of 94.7% overall, a 10 percentage point improvement over the FST baseline. This process demonstrates the feasibility of bootstrapping a neural morph analyzer from minimal resources.

Coupling Distant Annotation and Adversarial Training for Cross-Domain Chinese Word Segmentation [Website][PDF]
Ning Ding, Dingkun Long, Guangwei Xu, Muhua Zhu, Pengjun Xie, Xiaobin Wang, and Haitao Zheng
9:00–10:00

Fully supervised neural approaches have achieved significant progress in the task of Chinese word segmentation (CWS). Nevertheless, the performance of supervised models always drops gravely if the domain shifts due to the distribution gap across domains and the out of vocabulary (OOV) problem. In order to simultaneously alleviate the issues, this paper intuitively couples distant annotation and adversarial training for cross-domain CWS. 1) We rethink the essence of "Chinese words" and design an automatic distant annotation mechanism, which does not need any supervision or pre-defined dictionaries on the target domain. The method could effectively explore domain-specific words and distantly annotate the raw texts for the target domain. 2) We further develop a sentence-level adversarial training procedure to perform noise reduction and maximum utilization of the source domain information. Experiments on multiple real-world datasets across various domains show the superiority and robustness of our model, significantly outperforming previous state-of-the-arts cross-domain CWS methods.

Modeling Morphological Typology for Unsupervised Learning of Language Morphology [Website][PDF]
Hongzhi Xu, Jordan Kodner, Mitchell Marcus, and Charles Yang 9:00–10:00

This paper describes a language-independent model for fully unsupervised morphological analysis that exploits a universal framework leveraging morphological typology. By modeling morphological processes including suffixation, prefixation, infixation, and full and partial reduplication with constrained stem change rules, our system effectively constrains the search space and offers a wide coverage in terms of morphological typology. The system is tested on nine typologically and genetically diverse languages, and shows superior performance over leading systems. We also investigate the effect of an oracle that provides only a handful of bits per language to signal morphological type.

Predicting Declension Class from Form and Meaning [Website][PDF]
Adina Williams, Tiago Pimentel, Hagen Blix, Arya D. McCarthy, Eleanor Chodroff, and Ryan Cotterell
9:00–10:00

The noun lexica of many natural languages are divided into several declension classes with characteristic morphological properties. Class membership is far from deterministic, but the phonological form of a noun and/or its meaning can often provide imperfect clues. Here, we investigate the strength of those clues. More specifically, we operationalize this by measuring how much information, in bits, we can glean about declension class from knowing the form and/or meaning of nouns. We know that form and meaning are often also indicative of grammatical gender—which, as we quantitatively verify, can itself share information with declension class—so we also control for gender. We find for two Indo-European languages (Czech and German) that form and meaning respectively share significant amounts of information with class (and contribute additional information above and beyond gender). The three-way interaction between class, form, and meaning (given gender) is also significant. Our study is important for two reasons: First, we introduce a new method that provides additional quantitative support for a classic linguistic finding that form and meaning are relevant for the classification of nouns into declensions. Secondly, we show not only that individual declensions classes vary in the strength of their clues within a language, but also that these variations themselves vary across languages.

Predicting the Growth of Morphological Families from Social and Linguistic Factors [Website][PDF]
Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze 9:00–10:00

We present the first study that examines the evolution of morphological families, i.e., sets of morphologically related words such as “trump”, “antitrumpism”, and “detrumpify”, in social media. We introduce the novel task of Morphological Family Expansion Prediction (MFEP) as predicting the increase in the size of a morphological family. We create a ten-year Reddit corpus as a benchmark for MFEP and evaluate a number of baselines on this benchmark. Our experiments demonstrate very good performance on MFEP.

Semi-supervised Contextual Historical Text Normalization [Website] [PDF]
Peter Makarov and Simon Clematide 9:00–10:00

Historical text normalization, the task of mapping historical word forms to their modern counterparts, has recently attracted a lot of interest (Bollmann, 2019; Tang et al., 2018; Lusetti et al., 2018; Bollmann et al., 2018;Robertson and Goldwater, 2018; Bollmannet al., 2017; Korchagina, 2017). Yet, virtually all approaches suffer from the two limitations: 1) They consider a fully supervised setup, often with impractically large manually normalized datasets; 2) Normalization happens on words in isolation. By utilizing a simple generative normalization model and obtaining powerful contextualization from the target-side language model, we train accurate models with unlabeled historical data. In realistic training scenarios, our approach often leads to reduction in manually normalized data at the same accuracy levels.

Session 12B: Question Answering-10

ClarQ: A large-scale and diverse dataset for Clarification Question Generation [Website][PDF]
Vaibhav Kumar and Alan W Black 9:00–10:00

Question answering and conversational systems are often baffled and need help clarifying certain ambiguities. However, limitations of existing datasets hinder the development of large-scale models capable of generating and utilising clarification questions. In order to overcome these limitations, we devise a novel bootstrapping framework (based on self-supervision) that assists in the creation of a diverse, large-scale dataset of clarification questions based on post-comment tuples extracted from stackexchange. The framework utilises a neural network based architecture for classifying clarification questions. It is a two-step method where the first aims to increase the precision of the classifier and second aims to increase its recall. We quantitatively demonstrate the utility of the newly created dataset by applying it to the downstream task of question-answering. The final dataset, ClarQ, consists of ~2M examples distributed across 173 domains of stackexchange. We release this dataset in order to foster research into the field of clarification question generation with the larger goal of enhancing dialog and question answering systems.

DoQA - Accessing Domain-Specific FAQs via Conversational QA [Website][PDF]
Jon Ander Campos, Arantxa Otegi, Aitor Soroa, Jan Deriu, Mark Cieliebak, and Eneko Agirre 9:00–10:00

The goal of this work is to build conversational Question Answering (QA) interfaces for the large body of domain-specific information available in FAQ sites. We present DoQA, a dataset with 2,437 dialogues and 10,917 QA pairs. The dialogues are collected from three Stack Exchange sites using the Wizard of Oz method with crowdsourcing. Compared to previous work, DoQA comprises well-defined information needs, leading to more coherent and natural conversations with less factoid questions and is multi-domain. In addition, we introduce a more realistic information retrieval (IR) scenario where the system needs to find the answer in any of the FAQ documents. The results of an existing, strong, system show that, thanks to transfer learning from a Wikipedia QA dataset and fine tuning on a single FAQ domain, it is possible to build high quality conversational QA systems for FAQs without in-domain training data. The good results carry over into the more challenging IR scenario. In both cases, there is still ample room for improvement, as indicated by the higher human upperbound.

Harvesting and Refining Question-Answer Pairs for Unsupervised QA [Website][PDF]
Zhongli Li, Wenhui Wang, Li Dong, Furu Wei, and Ke Xu 9:00–10:00

Question Answering (QA) has shown great success thanks to the availability of large-scale datasets and the effectiveness of neural models. Recent research works have attempted to extend these successes to the settings with few or no labeled data available. In this work, we introduce two approaches to improve unsupervised QA. First, we harvest lexically and syntactically divergent questions from Wikipedia to automatically construct a corpus of question-answer pairs (named as RefQA). Second, we take advantage of the QA model to extract more appropriate answers, which iteratively refines data over RefQA. We conduct experiments on SQuAD 1.1, and NewsQA by fine-tuning BERT without access to manually annotated data. Our approach outperforms previous unsupervised approaches by a large margin, and is competitive with early supervised models. We also show the effectiveness of our approach in the few-shot learning setting.

MLQA: Evaluating Cross-lingual Extractive Question Answering [Website][PDF]
Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk 9:00–10:00

Question answering (QA) models have shown rapid progress enabled by the availability of large, high-quality benchmark datasets. Such annotated datasets are difficult and costly to collect, and rarely exist in languages other than English, making building QA systems that work well in other languages challenging. In order to develop such systems, it is crucial to invest in high quality multilingual evaluation benchmarks to measure progress. We present MLQA, a multi-way aligned extractive QA evaluation benchmark intended to spur research in this area. MLQA contains QA instances in 7 languages, English, Arabic, German, Spanish, Hindi, Vietnamese and Simplified Chinese. MLQA has over 12K instances in English and 5K in each other language, with each instance parallel between 4 languages on average. We evaluate state-of-the-art cross-lingual models and machine-translation-based baselines on MLQA. In all cases, transfer results are shown to be significantly behind training-language performance.

Multi-source Meta Transfer for Low Resource Multiple-Choice Question Answering [Website][PDF]
Ming Yan, Hao Zhang, Di Jin, and Joey Tianyi Zhou 9:00–10:00

Multiple-choice question answering (MCQA) is one of the most challenging tasks in machine reading comprehension since it requires more advanced reading comprehension skills such as logical reasoning, summarization, and arithmetic operations. Unfortunately, most existing MCQA datasets are small in size, which increases the difficulty of model learning and generalization. To address this challenge, we propose a multi-source meta transfer (MMT) for low-resource MCQA. In this framework, we first extend meta learning by incorporating multiple training sources to learn a generalized feature representation across domains. To bridge the distribution gap between training sources and the target, we further introduce the meta transfer that can be integrated into the multi-source meta training. More importantly, the proposed MMT is independent of backbone language models. Extensive experiments demonstrate the superiority of MMT over state-of-the-arts, and continuous improvements can be achieved on different backbone networks on both supervised and unsupervised domain adaptation settings.

RikiNet: Reading Wikipedia Pages for Natural Question Answering [Website][PDF]
Dayiheng Liu, Yeyun Gong, Jie Fu, Yu Yan, Jiusheng Chen, Daxin Jiang, Jiancheng Lv, and Nan Duan
9:00–10:00

Reading long documents to answer open-domain questions remains challenging in natural language understanding. In this paper, we introduce a new model, called RikiNet, which reads Wikipedia pages for natural question answering. RikiNet contains a dynamic paragraph dual-attention reader and a multi-level cascaded answer predictor. The reader dynamically represents the document and question by utilizing a set of complementary attention mechanisms. The representations are then fed into the predictor to obtain the span of the short answer, the paragraph of the long answer, and the answer type in a cascaded manner. On the Natural Questions (NQ) dataset, a single RikiNet achieves 74.3 F1 and 57.9 F1 on long-answer and short-answer tasks. To our best knowledge, it is the first single model that outperforms the single human performance. Furthermore, an ensemble RikiNet obtains 76.1 F1 and 61.3 F1 on long-answer and short-answer tasks, achieving the best performance on the official NQ leaderboard.

Session 12B Semantics: Textual Inference and Other Areas of Semantics-4

Do Neural Models Learn Systematicity of Monotonicity Inference in Natural Language? [Website] [PDF]
Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, and Kentaro Inui 9:00–10:00

Despite the success of language models using neural networks, it remains unclear to what extent neural models have the generalization ability to perform inferences. In this paper, we introduce a method for evaluating whether neural models can learn systematicity of monotonicity inference in natural language, namely, the regularity for performing arbitrary inferences with generalization on composition. We consider four aspects of monotonicity inferences and test whether the models can systematically interpret lexical and logical phenomena on different training/test splits. A series of experiments show that three neural models systematically draw inferences on unseen combinations of lexical and logical phenomena when the syntactic structures of the sentences are similar between the training and test sets. However, the performance of the models significantly decreases when the structures are slightly changed in the test set while retaining all vocabularies and constituents already appearing in the training set. This indicates that the generalization ability of neural models is limited to cases where the syntactic structures are nearly the same as those in the training set.

Fine-grained Fact Verification with Kernel Graph Attention Network [Website] [PDF]
Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu 9:00–10:00

Fact Verification requires fine-grained natural language inference capability that finds subtle clues to identify the syntactical and semantically correct but not well-supported claims. This paper presents Kernel Graph Attention Network (KGAT), which conducts more fine-grained fact verification with kernel-based attentions. Given a claim and a set of potential evidence sentences that form an evidence graph, KGAT introduces node kernels, which better measure the importance of the evidence node, and edge kernels, which conduct fine-grained evidence propagation in the graph, into Graph Attention Networks for more accurate fact verification. KGAT achieves a 70.38% FEVER score and significantly outperforms existing fact verification models on FEVER, a large-scale benchmark for fact verification. Our analyses illustrate that, compared to dot-product attentions, the kernel-based attention concentrates more on relevant evidence sentences and meaningful clues in the evidence graph, which is the main source of KGAT's effectiveness. All source codes of this work are available at https://github.com/thunlp/KernelGAT.

Generating Fact Checking Explanations [Website] [PDF]
Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein 9:00–10:00

Most existing work on automated fact checking is concerned with predicting the veracity of claims based on metadata, social network spread, language used in claims, and, more recently, evidence supporting or denying claims. A crucial piece of the puzzle that is still missing is to understand how to automate the most elaborate part of the process – generating justifications for verdicts on claims. This paper provides the first study of how these explanations can be generated automatically based on available claim context, and how this task can be modelled jointly with veracity prediction. Our results indicate that optimising both objectives at the same time, rather than training them separately, improves the performance of a fact checking system. The results of a manual evaluation further suggest that the informativeness, coverage and overall quality of the generated explanations are also improved in the multi-task model.

How to Ask Good Questions? Try to Leverage Paraphrases [Website] [PDF]
Xin Jia, Wenjie Zhou, Xu SUN, and Yunfang Wu 9:00–10:00

Given a sentence and its relevant answer, how to ask good questions is a challenging task, which has many real applications. Inspired by human's paraphrasing capability to ask questions of the same meaning but with diverse expressions, we propose to incorporate paraphrase knowledge into question generation(QG) to generate human-like questions. Specifically, we present a two-hand hybrid model leveraging a self-built paraphrase resource, which is automatically conducted by a simple back-translation method. On the one hand, we conduct multi-task learning with sentence-level paraphrase generation (PG) as an auxiliary task to supplement paraphrase knowledge to the task-share encoder. On the other hand, we adopt a new loss function for diversity training to introduce more question patterns to QG. Extensive experimental results show that our proposed model obtains obvious performance gain over several strong baselines, and further human evaluation validates that our model can ask questions of high quality by leveraging paraphrase knowledge.

Premise Selection in Natural Language Mathematical Texts [Website] [PDF]
Deborah Ferreira and André Freitas 9:00–10:00

The discovery of supporting evidence for addressing complex mathematical problems is a semantically challenging task, which is still unexplored in the field of natural language processing for mathematical text. The natural language premise selection task consists in using conjectures written in both natural language and mathematical formulae to recommend premises that most likely will be useful to prove a particular statement. We propose an approach to solve this task as a link prediction problem, using Deep Convolutional Graph Neural Networks. This paper also analyses how different baselines perform in this task and shows that a graph structure can provide higher F1-score, especially when considering multi-hop premise selection.

Session 12B: Student Research Workshop

Self-Attention is Not Only a Weight: Analyzing BERT with Vector Norms [Website]
Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui 9:00–10:00

Self-attention modules are essential building blocks of Transformer-based language models and hence are the subject of a large number of studies aiming to discover which linguistic capabilities these models possess (Rogers et al., 2020). Such studies are commonly conducted by analyzing correlations of attention weights with specific linguistic phenomena. In this paper, we show that attention weights alone are only one of two factors determining the output of self-attention modules and propose to incorporate the other factor, namely the norm of the transformed input vectors, into the analysis, as well. Our analysis of self-attention modules in BERT (Devlin et al., 2019) shows that the proposed method produces insights that better agree with linguistic intuitions than an analysis based on attention-weights alone. Our analysis further reveals that BERT controls the amount of the contribution from frequent informative and less informative tokens not by attention weights but via vector norms.

Transferring Monolingual Model to Low-Resource Language: The Case of Tigrinya [Website]
Abrhalei Frezghi Tela, Abraham Woubie Zewoudie, and Ville Hautamäki 9:00–10:00

In recent years, transformer models have achieved great success in natural language processing tasks. Most of the current state-of-the-art NLP results are achieved by using monolingual transformer models, where the model is pre-trained using a single language unlabelled text corpus. Then, the model is fine-tuned to the specific downstream task. However, the cost of pre-training a new transformer model is high for most languages. In this work, we propose a novel transfer learning method to adopt a strong source language model, trained from a large monolingual corpus to a low-resource language. Thus, using XLNet language model, we demonstrate competitive performance with mBERT and a pre-trained target language model on the Cross-lingual Sentiment (CLS) dataset and on a new sentiment analysis dataset for low-resourced language Tigrinya. With only 10k examples of the given Tigrinya sentiment analysis dataset, English XLNet has achieved 78.88% F1-Score outperforming BERT and mBERT by 10% and 7%, respectively. More interestingly, fine-tuning (English) XLNet model on the CLS dataset has promising results compared to mBERT and even outperformed mBERT for one dataset of the Japanese language.

Adaptive Transformers for Learning Multimodal Representations [Website][PDF]
Prajjwal Bhargava 9:00–10:00

The usage of transformers has grown from learning about language semantics to forming meaningful visiolinguistic representations. These architectures are often over-parametrized, requiring large amounts of computation. In this work, we extend adaptive approaches to learn more about model interpretability and computational efficiency. Specifically, we study attention spans, sparse, and structured dropout methods to help understand how their attention mechanism extends for vision and language tasks. We further show that these approaches can help us learn more about how the network perceives the complexity of input sequences, sparsity preferences for different modalities, and other related phenomena.

Session 12B: Theme-4

A Call for More Rigor in Unsupervised Cross-lingual Learning [Website][PDF]
Mikel Artetxe, Sebastian Ruder, Dani Yogatama, Gorka Labaka, and Eneko Agirre 9:00–10:00

We review motivations, definition, approaches, and methodology for unsupervised cross-lingual learning and call for a more rigorous position in each of them. An existing rationale for such research is based on the lack of parallel data for many of the world's languages. However, we argue that a scenario without any parallel data and abundant monolingual data is unrealistic in practice. We also discuss different training signals that have been used in previous work, which depart from the pure unsupervised setting. We then describe common methodological issues in tuning and evaluation of unsupervised cross-lingual models and present best practices. Finally, we provide a unified outlook for different types of research in this area (i.e., cross-lingual word embeddings, deep multilingual pretraining, and unsupervised machine translation) and argue for comparable evaluation of these models.

A Tale of a Probe and a Parser [Website][PDF]
Rowan Hall Maudslay, Josef Valvoda, Tiago Pimentel, Adina Williams, and Ryan Cotterell 9:00–10:00

Measuring what linguistic information is encoded in neural models of language has become popular in NLP. Researchers approach this enterprise by training "probes"—supervised models designed to extract linguistic structure from another model's output. One such probe is the structural probe (Hewitt and Manning, 2019), designed to quantify the extent to which syntactic information is encoded in contextualised word representations. The structural probe has a novel design, unattested in the parsing literature, the precise benefit of which is not immediately obvious. To explore whether syntactic probes would do better to make use of existing techniques, we compare the structural probe to a more traditional parser with an identical lightweight parameterisation. The parser outperforms structural probe on UUAS in seven of nine analysed languages, often by a substantial amount (e.g. by 11.1 points in English). Under a second less common metric, however, there is the opposite trend—the structural probe outperforms the parser. This begs the question: which metric should we prefer?

From SPMRL to NMRL: What Did We Learn (and Unlearn) in a Decade of Parsing Morphologically-Rich Languages (MRLs)? [Website][PDF]
Reut Tsarfaty, Dan Bareket, Stav Klein, and Amit Seker 9:00–10:00

It has been exactly a decade since the first establishment of SPMRL, a research initiative unifying multiple research efforts to address the peculiar challenges of Statistical Parsing for Morphologically-Rich Languages (MRLs). Here we reflect on parsing MRLs in that decade, highlight the solutions and lessons learned for the architectural, modeling and lexical challenges in the pre-neural era, and argue that similar challenges re-emerge in neural architectures for MRLs. We then aim to offer a climax, suggesting that incorporating symbolic ideas proposed in SPMRL terms into nowadays neural architectures has the potential to push NLP for MRLs to a new level. We sketch a strategies for designing Neural Models for MRLs (NMRL), and showcase preliminary support for these strategies via investigating the task of multi-tagging in Hebrew, a morphologically-rich, high-fusion, language.

Speech Translation and the End-to-End Promise: Taking Stock of Where We Are [Website][PDF]
Matthias Sperber and Matthias Paulik 9:00–10:00

Over its three decade history, speech translation has experienced several shifts in its primary research themes; moving from loosely coupled cascades of speech recognition and machine translation, to exploring questions of tight coupling, and finally to end-to-end models that have recently attracted much attention. This paper provides a brief survey of these developments, along with a discussion of the main challenges of traditional approaches which stem from committing to intermediate representations from the speech recognizer, and from training cascaded models separately towards different objectives. Recent end-to-end modeling techniques promise a principled way of overcoming these issues by allowing joint training of all model components and removing the need for explicit intermediate representations. However, a closer look reveals that many end-to-end models fall short of solving these issues, due to compromises made to address data scarcity. This paper provides a unifying categorization and nomenclature that covers both traditional and recent approaches and that may help researchers by highlighting both trade-offs and open research questions.

The State and Fate of Linguistic Diversity and Inclusion in the NLP World [Website][PDF]
Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury 9:00–10:00

Language technologies contribute to promoting multilingualism and linguistic diversity around the world. However, only a very small number of the over 7000 languages of the world are represented in the rapidly evolving language technologies and applications. In this paper we look at the relation between the types of languages, resources, and their representation in NLP conferences to understand the trajectory that different languages have followed over time. Our quantitative investigation underlines the disparity between languages, especially in terms of their resources, and calls into question the "language agnostic" status of current models and systems. Through this paper, we attempt to convince the ACL community to prioritise the resolution of the predicaments highlighted here, so that no language is left behind.

What Question Answering can Learn from Trivia Nerds [Website][PDF]
Jordan Boyd-Graber and Benjamin Börschinger 9:00–10:00

In addition to the traditional task of machines answering questions, question answering (QA) research creates interesting, challenging questions that help systems how to answer questions and reveal the best systems. We argue that creating a QA dataset—and the ubiquitous leaderboard that goes with it—closely resembles running a trivia tourna-

ment: you write questions, have agents (either humans or machines) answer the questions, and declare a winner. However, the research community has ignored the hard-learned lessons from decades of the trivia community creating vibrant, fair, and effective question answering competitions. After detailing problems with existing QA datasets, we outline the key lessons—removing ambiguity, discriminating skill, and adjudicating disputes—that can transfer to QA research and how they might be implemented.

What are the Goals of Distributional Semantics? [Website] [PDF]
Guy Emerson 9:00–10:00

Distributional semantic models have become a mainstay in NLP, providing useful features for downstream tasks. However, assessing long-term progress requires explicit long-term goals. In this paper, I take a broad linguistic perspective, looking at how well current models can deal with various semantic challenges. Given stark differences between models proposed in different subfields, a broad perspective is needed to see how we could integrate them. I conclude that, while linguistic insights can guide the design of model architectures, future progress will require balancing the often conflicting demands of linguistic expressiveness and computational tractability.

Demo Session 2C

Time: *9:30–10:15*

ADVISER: A Toolkit for Developing Multi-modal, Multi-domain and Socially-engaged Conversational Agents [Website][PDF]

Chia-Yu Li, Daniel Ortega, Dirk Väth, Florian Lux, Lindsey Vanderlyn, Maximilian Schmidt, Michael Neumann, Moritz Völkel, Pavel Denisov, Sabrina Jenne, Zorica Kacarevic, and Ngoc Thang Vu

We present ADVISER - an open-source, multi-domain dialog system toolkit that enables the development of multi-modal (incorporating speech, text and vision), socially-engaged (e.g. emotion recognition, engagement level prediction and backchanneling) conversational agents. The final Python-based implementation of our toolkit is flexible, easy to use, and easy to extend not only for technically experienced users, such as machine learning researchers, but also for less technically experienced users, such as linguists or cognitive scientists, thereby providing a flexible platform for collaborative research.

Demo Session 3A

Time: *12:00–12:45*

Torch-Struct: Deep Structured Prediction Library [Website][PDF]
Alexander Rush

The literature on structured prediction for NLP describes a rich collection of distributions and algorithms over sequences, segmentations, alignments, and trees; however, these algorithms are difficult to utilize in deep learning frameworks. We introduce Torch-Struct, a library for structured prediction designed to take advantage of and integrate with vectorized, auto-differentiation based frameworks. Torch-Struct includes a broad collection of probabilistic structures accessed through a simple and flexible distribution-based API that connects to any deep learning model. The library utilizes batched, vectorized operations and exploits auto-differentiation to produce readable, fast, and testable code. Internally, we also include a number of general-purpose optimizations to provide cross-algorithm efficiency. Experiments show significant performance gains over fast baselines and case-studies demonstrate the benefits of the library. Torch-Struct is available at https://github.com/harvardnlp/pytorch-struct.

Conversation Learner - A Machine Teaching Tool for Building Dialog Managers for Task-Oriented Dialog Systems [Website][PDF]
Swadheen Shukla, Lars Liden, Shahin Shayandeh, Eslam Kamal, Jinchao Li, Matt Mazzola, Thomas Park, Baolin Peng, and Jianfeng Gao

Traditionally, industry solutions for building a task-oriented dialog system have relied on helping dialog authors define rule-based dialog managers, represented as dialog flows. While dialog flows are intuitively interpretable and good for simple scenarios, they fall short of performance in terms of the flexibility needed to handle complex dialogs. On the other hand, purely machine-learned models can handle complex dialogs, but they are considered to be black boxes and require large amounts of training data. In this demonstration, we showcase Conversation Learner, a machine teaching tool for building dialog managers. It combines the best of both approaches by enabling dialog authors to create a dialog flow using familiar tools, converting the dialog flow into a parametric model (e.g., neural networks), and allowing dialog authors to improve the dialog manager (i.e., the parametric model) over time by leveraging user-system dialog logs as training data through a machine teaching interface.

ESPnet-ST: All-in-One Speech Translation Toolkit [Website][PDF]
Hirofumi Inaguma, Shun Kiyono, Kevin Duh, Shigeki Karita, Nelson Yalta, Tomoki Hayashi, and Shinji Watanabe

We present ESPnet-ST, which is designed for the quick development of speech-to-speech translation systems in a single framework. ESPnet-ST is a new project inside end-to-end speech processing toolkit, ESPnet, which integrates or newly implements automatic speech recognition, machine translation, and text-to-speech functions for speech translation. We provide all-in-one recipes including data pre-processing, feature extraction, training, and decoding pipelines for a wide range of benchmark datasets. Our reproducible results can match or even outperform the current state-of-the-art performances; these pre-trained models are downloadable. The toolkit is publicly available at https://github.com/espnet/espnet.

Session 13A Overview – Wednesday, July 8, 2020 12:00–13:00

Track A *Generation-12* Abstracts	Exploring Contextual Word-level Style Relevance for Unsupervised Style Transfer *Zhou, Chen, Liu, Xiao, Su, Guo, and Wu* [Website][PDF]	Heterogeneous Graph Transformer for Graph-to-Sequence Learning *Yao, Wang, and Wan* [Website][PDF]	Improving Image Captioning with Better Use of Caption *Shi, Zhou, Qiu, and Zhu* [Website][PDF]	Neural Data-to-Text Generation via Jointly Learning the Segmentation and Correspondence *Shen, Chang, Su, Niu, and Klakow* [Website][PDF]	Shape of Synth to Come: Why We Should Use Synthetic Data for English Surface Realization *Elder, Burke, O'Connor, and Foster* [Website][PDF]
	[TACL] Syntax-guided Controlled Generation of Paraphrases *Kumar, Ahuja, Vadapalli, and Talukdar* [Website][PDF]	Toward Better Storylines with Sentence-Level Language Models *Ippolito, Grangier, Eck, and Callison-Burch* [Website][PDF]			
Track B *Information Extraction-7* Abstracts	A Two-Step Approach for Implicit Event Argument Detection *Zhang, Kong, Liu, Ma, and Hovy* [Website][PDF]	A Unified MRC Framework for Named Entity Recognition *Li, Feng, Meng, Han, Wu, and Li* [Website][PDF]	Continual Relation Learning via Episodic Memory Activation and Reconsolidation *Han, Dai, Gao, Lin, Liu, Li, Sun, and Zhou* [Website][PDF]	[TACL] Improving Candidate Generation for Low-resource Cross-lingual Entity Linking *Zhou, Rijhwani, Wieting, Carbonell, and Neubig* [Website][PDF]	Improving Entity Linking through Semantic Reinforced Entity Embeddings *Hou, Wang, He, and Zhou* [Website][PDF]
	Machine Reading of Historical Events *Honovich, Torroba Hennigen, Abend, and Cohen* [Website][PDF]	Relation Extraction with Explanation *Shahbazi, Fern, Ghaeini, and Tadepalli* [Website][PDF]	Revisiting Unsupervised Relation Extraction *Tran, Le, and Ananiadou* [Website][PDF]	SciREX: A Challenge Dataset for Document-Level Information Extraction *Jain, Zuylen, Hajishirzi, and Beltagy* [Website][PDF]	Single-/Multi-Source Cross-Lingual NER via Teacher-Student Learning on Unlabeled Data in Target Language *Wu, Lin, Karlsson, LOU, and Huang* [Website][PDF]
	Synchronous Double-channel Recurrent Network for Aspect-Opinion Pair Extraction *Chen, Liu, Wang, Zhang, and Chi* [Website][PDF]				
Track C *Machine Learning for NLP-15* Abstracts	Contrastive Self-Supervised Learning for Commonsense Reasoning *Klein and Nabi* [Website][PDF]	Do Transformers Need Deep Long-Range Memory? *Rae and Razavi* [Website][PDF]	Effective Estimation of Deep Generative Language Models *Pelsmaeker and Aziz* [Website][PDF]	Exploiting Syntactic Structure for Better Language Modeling: A Syntactic Distance Approach *Du, Lin, Shen, O'Donnell, Bengio, and Zhang* [Website][PDF]	Highway Transformer: Self-Gating Enhanced Self-Attentive Networks *Chai, Jin, and Hou* [Website][PDF]

	Improving Disentangled Text Representation Learning with Information-Theoretic Guidance *Cheng, Min, Shen, Malon, Zhang, Li, and Carin* [Website][PDF]	Low-Dimensional Hyperbolic Knowledge Graph Embeddings *Chami, Wolf, Juan, Sala, Ravi, and Ré* [Website][PDF]	Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection *Ravfogel, Elazar, Gonen, Twiton, and Goldberg* [Website][PDF]		
Track D *NLP Applications-10* Abstracts	Closing the Gap: Joint De-Identification and Concept Extraction in the Clinical Domain *Lange, Adel, and Strötgen* [Website][PDF]	CorefQA: Coreference Resolution as Query-based Span Prediction *Wu, Wang, Yuan, Wu, and Li* [Website][PDF]	From Zero to Hero: Human-In-The-Loop Entity Linking in Low Resource Domains *Klie, Eckart de Castilho, and Gurevych* [Website][PDF]	Language to Network: Conditional Parameter Adaptation with Natural Language Descriptions *Jin, Liu, Yan, Eichenberger, and Morency* [Website][PDF]	Paraphrase Generation by Learning How to Edit from Samples *Kazemnejad, Salehi, and Soleymani Baghshah* [Website][PDF]
	Understanding Advertisements with BERT *Kalra, Kurma, Vadakkeeveetil Sreelatha, Patwardhan, and Karande* [Website][PDF]				
Track E *Lexical-7* Abstracts	[CL] LESSLEX: Linking Multilingual Embeddings to SenSe Representations of Lexical Items *Colla, Mensa, and Radicioni* [Website][PDF]	Non-Linear Instance-Based Cross-Lingual Mapping for Non-Isomorphic Embedding Spaces *Glavaš and Vulić* [Website][PDF]			
Track F *Sentence Level-8* Abstracts	Cross-Lingual Semantic Role Labeling with High-Quality Translated Training Corpus *Fei, Zhang, and Ji* [Website][PDF]	FastBERT: a Self-distilling BERT with Adaptive Inference Time *Liu, Zhou, Wang, Zhao, Deng, and JU* [Website][PDF]	Good-Enough Compositional Data Augmentation *Andreas* [Website][PDF]	LogicalFactChecker: Leveraging Logical Operations for Fact Checking with Graph Module Network *Zhong, Tang, Feng, Duan, Zhou, Gong, Shou, Jiang, Wang, and Yin* [Website][PDF]	Parsing into Variable-in-situ Logico-Semantic Graphs *Chen and Sun* [Website][PDF]
	RAT-SQL: Relation-Aware Schema Encoding and Linking for Text-to-SQL Parsers *Wang, Shin, Liu, Polozov, and Richardson* [Website][PDF]	Semi-Supervised Semantic Dependency Parsing Using CRF Autoencoders *Jia, Ma, Cai, and Tu* [Website][PDF]	Sentence Meta-Embeddings for Unsupervised Semantic Textual Similarity *Poerner, Waltinger, and Schütze* [Website][PDF]	Transition-based Semantic Dependency Parsing with Pointer Networks *Fernández-González and Gómez-Rodríguez* [Website][PDF]	tBERT: Topic Models and BERT Joining Forces for Semantic Similarity Detection *Peinelt, Nguyen, and Liakata* [Website][PDF]
Track G *Textual Inference and Other Areas of Semantics-5* Abstracts	Curriculum Learning for Natural Language Understanding *Xu, Zhang, Mao, Wang, Xie, and Zhang* [Website][PDF]	Evidence-Aware Inferential Text Generation with Vector Quantised Variational AutoEncoder *Guo, Tang, Duan, Yin, Jiang, and Zhou* [Website][PDF]	Fine-grained Fact Verification with Kernel Graph Attention Network *Liu, Xiong, Sun, and Liu* [Website][PDF]	NeuInfer: Knowledge Inference on N-ary Facts *Guan, Jin, Guo, Wang, and Cheng* [Website][PDF]	Neural Graph Matching Networks for Chinese Short Text Matching *Chen, Zhao, Lyu, Jin, Chen, Zhu, and Yu* [Website][PDF]

	Premise Selection in Natural Language Mathematical Texts *Ferreira and Freitas* [Website][PDF]	Reasoning Over Semantic-Level Graph for Fact Checking *Zhong, Xu, Tang, Xu, Duan, Zhou, Wang, and Yin* [Website][PDF]	Temporal Common Sense Acquisition with Minimal Supervision *Zhou, Ning, Khashabi, and Roth* [Website][PDF]	The Sensitivity of Language Models and Humans to Winograd Schema Perturbations *Abdou, Ravishankar, Barrett, Belinkov, Elliott, and Søgaard* [Website][PDF]	
Track H *Sentiment Analysis, Stylistic Analysis, and Argument Mining-11* Abstracts	Conditional Augmentation for Aspect Term Extraction via Masked Sequence-to-Sequence Generation *Li, Chen, Quan, Ling, and Song* [Website][PDF]	Exploiting Personal Characteristics of Debaters for Predicting Persuasiveness *Al Khatib, Völske, Syed, Kolyada, and Stein* [Website][PDF]	Out of the Echo Chamber: Detecting Countering Debate Speeches *Orbach, Bilu, Toledo, Lahav, Jacovi, Aharonov, and Slonim* [Website][PDF]		
Track I *Student Research Workshop* Abstracts	Pre-training via Leveraging Assisting Languages for Neural Machine Translation *Song, Dabre, Mao, Cheng, Kurohashi, and Sumita* [Website][PDF]	A Simple and Effective Dependency Parser for Telugu *Nallani, Shrivastava, and Sharma* [Website][PDF]	Cross-Lingual Disaster-related Multi-label Tweet Classification with Manifold Mixup *Ray Chowdhury, Caragea, and Caragea* [Website][PDF]		

Session 13A Details

Session 13A: Generation-12

Exploring Contextual Word-level Style Relevance for Unsupervised Style Transfer [Website][PDF]
Chulun Zhou, Liangyu Chen, Jiachen Liu, Xinyan Xiao, Jinsong Su, Sheng Guo, and Hua Wu 12:00–13:00

Unsupervised style transfer aims to change the style of an input sentence while preserving its original content without using parallel training data. In current dominant approaches, owing to the lack of fine-grained control on the influence from the target style, they are unable to yield desirable output sentences. In this paper, we propose a novel attentional sequence-to-sequence (Seq2seq) model that dynamically exploits the relevance of each output word to the target style for unsupervised style transfer. Specifically, we first pretrain a style classifier, where the relevance of each input word to the original style can be quantified via layer-wise relevance propagation. In a denoising auto-encoding manner, we train an attentional Seq2seq model to reconstruct input sentences and repredict word-level previously-quantified style relevance simultaneously. In this way, this model is endowed with the ability to automatically predict the style relevance of each output word. Then, we equip the decoder of this model with a neural style component to exploit the predicted wordlevel style relevance for better style transfer. Particularly, we fine-tune this model using a carefully-designed objective function involving style transfer, style relevance consistency, content preservation and fluency modeling loss terms. Experimental results show that our proposed model achieves state-of-the-art performance in terms of both transfer accuracy and content preservation.

Heterogeneous Graph Transformer for Graph-to-Sequence Learning [Website][PDF]
Shaowei Yao, Tianming Wang, and Xiaojun Wan 12:00–13:00

The graph-to-sequence (Graph2Seq) learning aims to transduce graph-structured representations to word sequences for text generation. Recent studies propose various models to encode graph structure. However, most previous works ignore the indirect relations between distance nodes, or treat indirect relations and direct relations in the same way. In this paper, we propose the Heterogeneous Graph Transformer to independently model the different relations in the individual subgraphs of the original graph, including direct relations, indirect relations and multiple possible relations between nodes. Experimental results show that our model strongly outperforms the state of the art on all four standard benchmarks of AMR-to-text generation and syntax-based neural machine translation.

Improving Image Captioning with Better Use of Caption [Website][PDF]
Zhan Shi, Xu Zhou, Xipeng Qiu, and Xiaodan Zhu 12:00–13:00

Image captioning is a multimodal problem that has drawn extensive attention in both the natural language processing and computer vision community. In this paper, we present a novel image captioning architecture to better explore semantics available in captions and leverage that to enhance both image representation and caption generation. Our models first construct caption-guided visual relationship graphs that introduce beneficial inductive bias using weakly supervised multi-instance learning. The representation is then enhanced with neighbouring and contextual nodes with their textual and visual features. During generation, the model further incorporates visual relationships using multi-task learning for jointly predicting word and object/predicate tag sequences. We perform extensive experiments on the MSCOCO dataset, showing that the proposed framework significantly outperforms the baselines, resulting in the state-of-the-art performance under a wide range of evaluation metrics. The code of our paper has been made publicly available.

Neural Data-to-Text Generation via Jointly Learning the Segmentation and Correspondence [Website][PDF]
Xiaoyu Shen, Ernie Chang, Hui Su, Cheng Niu, and Dietrich Klakow 12:00–13:00

The neural attention model has achieved great success in data-to-text generation tasks. Though usually excelling at producing fluent text, it suffers from the problem of information missing, repetition and "hallucination". Due to the black-box nature of the neural attention architecture, avoiding these problems in a systematic way is non-trivial. To address this concern, we propose to explicitly segment target text into fragment units and align them with their data correspondences. The segmentation and correspondence are jointly learned as latent variables without any human annotations. We further impose a soft statistical constraint to regularize the segmental granularity. The resulting architecture maintains the same expressive power as neural attention models, while being able to generate fully interpretable outputs with several times less computational cost. On both E2E and WebNLG benchmarks, we show the proposed model consistently outperforms its neural attention counterparts.

Shape of Synth to Come: Why We Should Use Synthetic Data for English Surface Realization [Website][PDF]
Henry Elder, Robert Burke, Alexander O'Connor, and Jennifer Foster 12:00–13:00

The Surface Realization Shared Tasks of 2018 and 2019 were Natural Language Generation shared tasks with the goal of exploring approaches to surface realization from Universal-Dependency-like trees to surface strings for several languages. In the 2018 shared task there was very little difference in the absolute performance of systems trained with and without additional, synthetically created data, and a new rule prohibiting the use of synthetic data was introduced for the 2019 shared task. Contrary to the findings of the 2018 shared task, we show, in experiments on the English 2018 dataset, that the use of synthetic data can have a substantial positive effect – an improvement of almost 8 BLEU points for a previously state-of-the-art system. We analyse the effects of synthetic data, and we argue that its use should be

encouraged rather than prohibited so that future research efforts continue to explore systems that can take advantage of such data.

[TACL] Syntax-guided Controlled Generation of Paraphrases [Website][PDF]
Ashutosh Kumar, Kabir Ahuja, Raghuram Vadapalli, and Partha Talukdar 12:00–13:00

Given a sentence (e.g., "I like mangoes") and a constraint (e.g., negative sentiment), the goal of controlled text generation is to produce a sentence that adapts the input sentence to meet the requirements of the constraint (e.g., "I hate mangoes"). Going beyond such simple constraints, recent works have started exploring the incorporation of complex syntactic-guidance as constraints in the task of controlled paraphrase generation. In these methods, syntactic-guidance is sourced from a separate exemplar sentence. However, these prior works have only utilized limited syntactic information available in the parse tree of the exemplar sentence. We address this limitation in the paper and propose Syntax Guided Controlled Paraphraser (SGCP), an end-to-end framework for syntactic paraphrase generation. We find that S GCP can generate syntax-conforming sentences while not compromising on relevance. We perform extensive automated and human evaluations over multiple real-world datasets to demonstrate the efficacy of SGCP over state-of-the-art baselines. To drive future research, we have made SGCP's source code available.

Toward Better Storylines with Sentence-Level Language Models [Website][PDF]
Daphne Ippolito, David Grangier, Douglas Eck, and Chris Callison-Burch 12:00–13:00

We propose a sentence-level language model which selects the next sentence in a story from a finite set of fluent alternatives. Since it does not need to model fluency, the sentence-level language model can focus on longer range dependencies, which are crucial for multi-sentence coherence. Rather than dealing with individual words, our method treats the story so far as a list of pre-trained sentence embeddings and predicts an embedding for the next sentence, which is more efficient than predicting word embeddings. Notably this allows us to consider a large number of candidates for the next sentence during training. We demonstrate the effectiveness of our approach with state-of-the-art accuracy on the unsupervised Story Cloze task and with promising results on larger-scale next sentence prediction tasks.

Session 13A: Information Extraction-7

A Two-Step Approach for Implicit Event Argument Detection [Website][PDF]
Zhisong Zhang, Xiang Kong, Zhengzhong Liu, Xuezhe Ma, and Eduard Hovy 12:00–13:00

In this work, we explore the implicit event argument detection task, which studies event arguments beyond sentence boundaries. The addition of cross-sentence argument candidates imposes great challenges for modeling. To reduce the number of candidates, we adopt a two-step approach, decomposing the problem into two sub-problems: argument head-word detection and head-to-span expansion. Evaluated on the recent RAMS dataset (Ebner et al., 2020), our model achieves overall better performance than a strong sequence labeling baseline. We further provide detailed error analysis, presenting where the model mainly makes errors and indicating directions for future improvements. It remains a challenge to detect implicit arguments, calling for more future work of document-level modeling for this task.

A Unified MRC Framework for Named Entity Recognition [Website][PDF]
Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li 12:00–13:00

The task of named entity recognition (NER) is normally divided into nested NER and flat NER depending on whether named entities are nested or not.Models are usually separately developed for the two tasks, since sequence labeling models, the most widely used backbone for flat NER, are only able to assign a single label to a particular token, which is unsuitable for nested NER where a token may be assigned several labels. In this paper, we propose a unified framework that is capable of handling both flat and nested NER tasks. Instead of treating the task of NER as a sequence labeling problem, we propose to formulate it as a machine reading comprehension (MRC) task. For example, extracting entities with the PER label is formalized as extracting answer spans to the question "*which person is mentioned in the text*".This formulation naturally tackles the entity overlapping issue in nested NER: the extraction of two overlapping entities with different categories requires answering two independent questions. Additionally, since the query encodes informative prior knowledge, this strategy facilitates the process of entity extraction, leading to better performances for not only nested NER, but flat NER. We conduct experiments on both nested and flat NER datasets.Experiment results demonstrate the effectiveness of the proposed formulation. We are able to achieve a vast amount of performance boost over current SOTA models on nested NER datasets, i.e., +1.28, +2.55, +5.44, +6.37,respectively on ACE04, ACE05, GENIA and KBP17, along with SOTA results on flat NER datasets, i.e., +0.24, +1.95, +0.21, +1.49 respectively on English CoNLL 2003, English OntoNotes 5.0, Chinese MSRA and Chinese OntoNotes 4.0.

Continual Relation Learning via Episodic Memory Activation and Reconsolidation [Website][PDF]
Xu Han, Yi Dai, Tianyu Gao, Yankai Lin, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou 12:00–13:00

Continual relation learning aims to continually train a model on new data to learn incessantly emerging novel relations while avoiding catastrophically forgetting old relations. Some pioneering work has proved that storing a handful of historical relation examples in episodic memory and replaying them in subsequent training is an effective solution for such a challenging problem. However, these memory-based methods usually suffer from overfitting the few memorized examples of old relations, which may gradually cause inevitable confusion among existing relations. Inspired by the mechanism in human long-term memory formation, we introduce episodic memory activation and reconsolidation (EMAR) to continual relation learning. Every time neural models are activated to learn both new and memorized data, EMAR utilizes relation prototypes for memory reconsolidation exercise to keep a stable understanding of old relations. The experimental results show that EMAR could get rid of catastrophically forgetting old relations and outperform the state-of-the-art continual learning models.

[TACL] Improving Candidate Generation for Low-resource Cross-lingual Entity Linking [Website][PDF]
Shuyan Zhou, Shruti Rijhwani, John Wieting, Jaime Carbonell, and Graham Neubig 12:00–13:00

Cross-lingual entity linking (XEL) is the task of finding referents in a target-language knowledge base (KB) for mentions extracted from source-language texts. The first step of (X)EL is candidate generation, which retrieves a list of plausible candidate entities from the target-language KB for each mention. Approaches based on resources from Wikipedia have proven successful in the realm of relatively high-resource languages (HRL), but these do not extend well to low-resource languages (LRL) with few, if any, Wikipedia pages. Recently, transfer learning methods have been shown to reduce the demand for resources in the LRL by utilizing resources in closely-related languages, but the performance still lags far behind their high-resource counterparts. In this paper, we first assess the problems faced by current entity candidate generation methods for low-resource XEL, then propose three improvements that (1) reduce the disconnect between entity mentions and KB entries, and (2) improve the robustness of the model to low-resource scenarios. The methods are simple, but effective: we experiment with our approach on seven XEL datasets and find that they yield an average gain of 16.9% in Top-30 gold candidate recall, compared to state-of-the-art baselines. Our improved model also yields an average gain of 7.9% in in-KB accuracy of end-to-end XEL.

Improving Entity Linking through Semantic Reinforced Entity Embeddings [Website][PDF]
Feng Hou, Ruili Wang, Jun He, and Yi Zhou 12:00–13:00

Entity embeddings, which represent different aspects of each entity with a single vector like word embeddings, are a key component of neural entity linking models. Existing entity embeddings are learned from canonical Wikipedia articles and local contexts surrounding target entities. Such entity embeddings are effective, but too distinctive for linking models to learn contextual commonality. We propose a simple yet effective method, FGS2EE, to inject fine-grained semantic information into entity embeddings to reduce the distinctiveness and facilitate the learning of contextual commonality. FGS2EE first uses the embeddings of semantic type words to generate semantic embeddings, and then combines them with existing entity embeddings through linear aggregation. Extensive experiments show the effec-

tiveness of such embeddings. Based on our entity embeddings, we achieved new sate-of-the-art performance on entity linking.

Machine Reading of Historical Events [Website][PDF]
Or Honovich, Lucas Torroba Hennigen, Omri Abend, and Shay B. Cohen 12:00–13:00

Machine reading is an ambitious goal in NLP that subsumes a wide range of text understanding capabilities. Within this broad framework, we address the task of machine reading the time of historical events, compile datasets for the task, and develop a model for tackling it. Given a brief textual description of an event, we show that good performance can be achieved by extracting relevant sentences from Wikipedia, and applying a combination of task-specific and general-purpose feature embeddings for the classification. Furthermore, we establish a link between the historical event ordering task and the event focus time task from the information retrieval literature, showing they also provide a challenging test case for machine reading algorithms.

Relation Extraction with Explanation [Website][PDF]
Hamed Shahbazi, Xiaoli Fern, Reza Ghaeini, and Prasad Tadepalli 12:00–13:00

Recent neural models for relation extraction with distant supervision alleviate the impact of irrelevant sentences in a bag by learning importance weights for the sentences. Efforts thus far have focused on improving extraction accuracy but little is known about their explanability. In this work we annotate a test set with ground-truth sentence-level explanations to evaluate the quality of explanations afforded by the relation extraction models. We demonstrate that replacing the entity mentions in the sentences with their fine-grained entity types not only enhances extraction accuracy but also improves explanation. We also propose to automatically generate "distractor" sentences to augment the bags and train the model to ignore the distractors. Evaluations on the widely used FB-NYT dataset show that our methods achieve new state-of-the-art accuracy while improving model explanability.

Revisiting Unsupervised Relation Extraction [Website][PDF]
Thy Thy Tran, Phong Le, and Sophia Ananiadou 12:00–13:00

Unsupervised relation extraction (URE) extracts relations between named entities from raw text without manually-labelled data and existing knowledge bases (KBs). URE methods can be categorised into generative and discriminative approaches, which rely either on hand-crafted features or surface form. However, we demonstrate that by using only named entities to induce relation types, we can outperform existing methods on two popular datasets. We conduct a comparison and evaluation of our findings with other URE techniques, to ascertain the important features in URE. We conclude that entity types provide a strong inductive bias for URE.

SciREX: A Challenge Dataset for Document-Level Information Extraction [Website][PDF]
Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy 12:00–13:00

Extracting information from full documents is an important problem in many domains, but most previous work focus on identifying relationships within a sentence or a paragraph. It is challenging to create a large-scale information extraction (IE) dataset at the document level since it requires an understanding of the whole document to annotate entities and their document-level relationships that usually span beyond sentences or even sections. In this paper, we introduce SciREX, a document level IE dataset that encompasses multiple IE tasks, including salient entity identification and document level N-ary relation identification from scientific articles. We annotate our dataset by integrating automatic and human annotations, leveraging existing scientific knowledge resources. We develop a neural model as a strong baseline that extends previous state-of-the-art IE models to document-level IE. Analyzing the model performance shows a significant gap between human performance and current baselines, inviting the community to use our dataset as a challenge to develop document-level IE models. Our data and code are publicly available at https://github.com/allenai/SciREX .

Single-/Multi-Source Cross-Lingual NER via Teacher-Student Learning on Unlabeled Data in Target Language [Website][PDF]
Qianhui Wu, Zijia Lin, Börje Karlsson, Jian-Guang LOU, and Biqing Huang 12:00–13:00

To better tackle the named entity recognition (NER) problem on languages with little/no labeled data, cross-lingual NER must effectively leverage knowledge learned from source languages with rich labeled data. Previous works on cross-lingual NER are mostly based on label projection with pairwise texts or direct model transfer. However, such methods either are not applicable if the labeled data in the source languages is unavailable, or do not leverage information contained in unlabeled data in the target language. In this paper, we propose a teacher-student learning method to address such limitations, where NER models in the source languages are used as teachers to train a student model on unlabeled data in the target language. The proposed method works for both single-source and multi-source cross-lingual NER. For the latter, we further propose a similarity measuring method to better weight the supervision from different teacher models. Extensive experiments for 3 target languages on benchmark datasets well demonstrate that our method outperforms existing state-of-the-art methods for both single-source and multi-source cross-lingual NER.

Synchronous Double-channel Recurrent Network for Aspect-Opinion Pair Extraction [Website][PDF]
Shaowei Chen, Jie Liu, Yu Wang, Wenzheng Zhang, and Ziming Chi 12:00–13:00

Opinion entity extraction is a fundamental task in fine-grained opinion mining. Related studies generally extract aspects and/or opinion expressions without recognizing the relations between them. However, the relations are crucial for downstream tasks, including sentiment classification, opinion summarization, etc. In this paper, we explore Aspect-Opinion Pair Extraction (AOPE) task, which aims at extracting aspects and opinion expressions in pairs. To deal with this task, we propose Synchronous Double-channel Recurrent Network (SDRN) mainly consisting of an opinion entity extraction unit, a relation detection unit, and a synchronization unit. The opinion entity extraction

unit and the relation detection unit are developed as two channels to extract opinion entities and relations simultaneously. Furthermore, within the synchronization unit, we design Entity Synchronization Mechanism (ESM) and Relation Synchronization Mechanism (RSM) to enhance the mutual benefit on the above two channels. To verify the performance of SDRN, we manually build three datasets based on SemEval 2014 and 2015 benchmarks. Extensive experiments demonstrate that SDRN achieves state-of-the-art performances.

Session 13A: Machine Learning for NLP-15

Contrastive Self-Supervised Learning for Commonsense Reasoning [Website][PDF]
Tassilo Klein and Moin Nabi 12:00–13:00

We propose a self-supervised method to solve Pronoun Disambiguation and Winograd Schema Challenge problems. Our approach exploits the characteristic structure of training corpora related to so-called "trigger" words, which are responsible for flipping the answer in pronoun disambiguation. We achieve such commonsense reasoning by constructing pair-wise contrastive auxiliary predictions. To this end, we leverage a mutual exclusive loss regularized by a contrastive margin. Our architecture is based on the recently introduced transformer networks, BERT, that exhibits strong performance on many NLP benchmarks. Empirical results show that our method alleviates the limitation of current supervised approaches for commonsense reasoning. This study opens up avenues for exploiting inexpensive self-supervision to achieve performance gain in commonsense reasoning tasks.

Do Transformers Need Deep Long-Range Memory? [Website][PDF]
Jack Rae and Ali Razavi 12:00–13:00

Deep attention models have advanced the modelling of sequential data across many domains. For language modelling in particular, the Transformer-XL — a Transformer augmented with a long-range memory of past activations — has been shown to be state-of-the-art across a variety of well-studied benchmarks. The Transformer-XL incorporates a long-range memory at every layer of the network, which renders its state to be thousands of times larger than RNN predecessors. However it is unclear whether this is necessary. We perform a set of interventions to show that comparable performance can be obtained with 6X fewer long range memories and better performance can be obtained by limiting the range of attention in lower layers of the network.

Effective Estimation of Deep Generative Language Models [Website][PDF]
Tom Pelsmaeker and Wilker Aziz 12:00–13:00

Advances in variational inference enable parameterisation of probabilistic models by deep neural networks. This combines the statistical transparency of the probabilistic modelling framework with the representational power of deep learning. Yet, due to a problem known as posterior collapse, it is difficult to estimate such models in the context of language modelling effectively. We concentrate on one such model, the variational auto-encoder, which we argue is an important building block in hierarchical probabilistic models of language. This paper contributes a sober view of the problem, a survey of techniques to address it, novel techniques, and extensions to the model. To establish a ranking of techniques, we perform a systematic comparison using Bayesian optimisation and find that many techniques perform reasonably similar, given enough resources. Still, a favourite can be named based on convenience. We also make several empirical observations and recommendations of best practices that should help researchers interested in this exciting field.

Exploiting Syntactic Structure for Better Language Modeling: A Syntactic Distance Approach [Website][PDF]
Wenyu Du, Zhouhan Lin, Yikang Shen, Timothy J. O'Donnell, Yoshua Bengio, and Yue Zhang 12:00–13:00

It is commonly believed that knowledge of syntactic structure should improve language modeling. However, effectively and computationally efficiently incorporating syntactic structure into neural language models has been a challenging topic. In this paper, we make use of a multi-task objective, i.e., the models simultaneously predict words as well as ground truth parse trees in a form called "syntactic distances", where information between these two separate objectives shares the same intermediate representation. Experimental results on the Penn Treebank and Chinese Treebank datasets show that when ground truth parse trees are provided as additional training signals, the model is able to achieve lower perplexity and induce trees with better quality.

Highway Transformer: Self-Gating Enhanced Self-Attentive Networks [Website][PDF]
Yekun Chai, Shuo Jin, and Xinwen Hou 12:00–13:00

Self-attention mechanisms have made striking state-of-the-art (SOTA) progress in various sequence learning tasks, standing on the multi-headed dot product attention by attending to all the global contexts at different locations. Through a pseudo information highway, we introduce a gated component self-dependency units (SDU) that incorporates LSTM-styled gating units to replenish internal semantic importance within the multi-dimensional latent space of individual representations. The subsidiary content-based SDU gates allow for the information flow of modulated latent embeddings through skipped connections, leading to a clear margin of convergence speed with gradient descent algorithms. We may unveil the role of gating mechanism to aid in the context-based Transformer modules, with hypothesizing that SDU gates, especially on shallow layers, could push it faster to step towards suboptimal points during the optimization process.

Improving Disentangled Text Representation Learning with Information-Theoretic Guidance [Website][PDF]
Pengyu Cheng, Martin Renqiang Min, Dinghan Shen, Christopher Malon, Yizhe Zhang, Yitong Li, and Lawrence Carin 12:00–13:00

Learning disentangled representations of natural language is essential for many NLP tasks, e.g., conditional text generation, style transfer, personalized dialogue systems, etc. Similar problems have been studied extensively for other forms of data, such as images and videos. However, the discrete nature of natural language makes the disentangling of textual representations more challenging (e.g., the manipulation over the data space cannot be easily achieved). Inspired by information theory, we propose a novel method that effectively manifests disentangled representations of text, without any supervision on semantics. A new mutual information upper bound is derived and leveraged to

measure dependence between style and content. By minimizing this upper bound, the proposed method induces style and content embeddings into two independent low-dimensional spaces. Experiments on both conditional text generation and text-style transfer demonstrate the high quality of our disentangled representation in terms of content and style preservation.

Low-Dimensional Hyperbolic Knowledge Graph Embeddings [Website] [PDF]
Ines Chami, Adva Wolf, Da-Cheng Juan, Frederic Sala, Sujith Ravi, and Christopher Ré 12:00–13:00

Knowledge graph (KG) embeddings learn low- dimensional representations of entities and relations to predict missing facts. KGs often exhibit hierarchical and logical patterns which must be preserved in the embedding space. For hierarchical data, hyperbolic embedding methods have shown promise for high-fidelity and parsimonious representations. However, existing hyperbolic embedding methods do not account for the rich logical patterns in KGs. In this work, we introduce a class of hyperbolic KG embedding models that simultaneously capture hierarchical and logical patterns. Our approach combines hyperbolic reflections and rotations with attention to model complex relational patterns. Experimental results on standard KG benchmarks show that our method improves over previous Euclidean- and hyperbolic-based efforts by up to 6.1% in mean reciprocal rank (MRR) in low dimensions. Furthermore, we observe that different geometric transformations capture different types of relations while attention- based transformations generalize to multiple relations. In high dimensions, our approach yields new state-of-the-art MRRs of 49.6% on WN18RR and 57.7% on YAGO3-10.

Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection [Website] [PDF]
Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg 12:00–13:00

The ability to control for the kinds of information encoded in neural representation has a variety of use cases, especially in light of the challenge of interpreting these models. We present Iterative Null-space Projection (INLP), a novel method for removing information from neural representations. Our method is based on repeated training of linear classifiers that predict a certain property we aim to remove, followed by projection of the representations on their null-space. By doing so, the classifiers become oblivious to that target property, making it hard to linearly separate the data according to it. While applicable for multiple uses, we evaluate our method on bias and fairness use-cases, and show that our method is able to mitigate bias in word embeddings, as well as to increase fairness in a setting of multi-class classification.

Session 13A: NLP Applications-10

Closing the Gap: Joint De-Identification and Concept Extraction in the Clinical Domain [Website][PDF]
Lukas Lange, Heike Adel, and Jannik Strötgen 12:00–13:00

Exploiting natural language processing in the clinical domain requires de-identification, i.e., anonymization of personal information in texts. However, current research considers de-identification and downstream tasks, such as concept extraction, only in isolation and does not study the effects of de-identification on other tasks. In this paper, we close this gap by reporting concept extraction performance on automatically anonymized data and investigating joint models for de-identification and concept extraction. In particular, we propose a stacked model with restricted access to privacy sensitive information and a multitask model. We set the new state of the art on benchmark datasets in English (96.1% F1 for de-identification and 88.9% F1 for concept extraction) and Spanish (91.4% F1 for concept extraction).

CorefQA: Coreference Resolution as Query-based Span Prediction [Website][PDF]
Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li 12:00–13:00

In this paper, we present CorefQA, an accurate and extensible approach for the coreference resolution task. We formulate the problem as a span prediction task, like in question answering: A query is generated for each candidate mention using its surrounding context, and a span prediction module is employed to extract the text spans of the coreferences within the document using the generated query. This formulation comes with the following key advantages: (1) The span prediction strategy provides the flexibility of retrieving mentions left out at the mention proposal stage; (2) In the question answering framework, encoding the mention and its context explicitly in a query makes it possible to have a deep and thorough examination of cues embedded in the context of coreferent mentions; and (3) A plethora of existing question answering datasets can be used for data augmentation to improve the model's generalization capability. Experiments demonstrate significant performance boost over previous models, with 83.1 (+3.5) F1 score on the CoNLL-2012 benchmark and 87.5 (+2.5) F1 score on the GAP benchmark.

From Zero to Hero: Human-In-The-Loop Entity Linking in Low Resource Domains [Website][PDF]
Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych 12:00–13:00

Entity linking (EL) is concerned with disambiguating entity mentions in a text against knowledge bases (KB). It is crucial in a considerable number of fields like humanities, technical writing and biomedical sciences to enrich texts with semantics and discover more knowledge. The use of EL in such domains requires handling noisy texts, low resource settings and domain-specific KBs. Existing approaches are mostly inappropriate for this, as they depend on training data. However, in the above scenario, there exists hardly annotated data, and it needs to be created from scratch. We therefore present a novel domain-agnostic Human-In-The-Loop annotation approach: we use recommenders that suggest potential concepts and adaptive candidate ranking, thereby speeding up the overall annotation process and making it less tedious for users. We evaluate our ranking approach in a simulation on difficult texts and show that it greatly outperforms a strong baseline in ranking accuracy. In a user study, the annotation speed improves by 35% compared to annotating without interactive support; users report that they strongly prefer our system. An open-source and ready-to-use implementation based on the text annotation platform INCEpTION (https://inception-project.github.io) is made available.

Language to Network: Conditional Parameter Adaptation with Natural Language Descriptions [Website][PDF]
Tian Jin, Zhun Liu, Shengjia Yan, Alexandre Eichenberger, and Louis-Philippe Morency 12:00–13:00

Transfer learning using ImageNet pre-trained models has been the de facto approach in a wide range of computer vision tasks. However, fine-tuning still requires task-specific training data. In this paper, we propose $\mathbf{N^3}$ (**N**eural **N**etworks from **N**atural Language) - a new paradigm of synthesizing task-specific neural networks from language descriptions and a generic pre-trained model. $\mathbf{N^3}$ leverages language descriptions to generate parameter adaptations as well as a new task-specific classification layer for a pre-trained neural network, effectively "fine-tuning" the network for a new task using only language descriptions as input. To the best of our knowledge, $\mathbf{N^3}$ is the first method to synthesize entire neural networks from natural language. Experimental results show that $\mathbf{N^3}$ can out-perform previous natural-language based zero-shot learning methods across 4 different zero-shot image classification benchmarks. We also demonstrate a simple method to help identify keywords in language descriptions leveraged by $\mathbf{N^3}$ when synthesizing model parameters.

Paraphrase Generation by Learning How to Edit from Samples [Website][PDF]
Amirhossein Kazemnejad, Mohammadreza Salehi, and Mahdieh Soleymani Baghshah 12:00–13:00

Neural sequence to sequence text generation has been proved to be a viable approach to paraphrase generation. Despite promising results, paraphrases generated by these models mostly suffer from lack of quality and diversity. To address these problems, we propose a novel retrieval-based method for paraphrase generation. Our model first retrieves a paraphrase pair similar to the input sentence from a pre-defined index. With its novel editor module, the model then paraphrases the input sequence by editing it using the extracted relations between the retrieved pair of sentences. In order to have fine-grained control over the editing process, our model uses the newly introduced concept of Micro Edit Vectors. It both extracts and exploits these vectors using the attention mechanism in the Transformer architecture. Experimental results show the superiority of our paraphrase generation method in terms of both automatic metrics, and human evaluation of relevance, grammaticality, and diversity of generated paraphrases.

Understanding Advertisements with BERT [Website] [PDF]

Kanika Kalra, Bhargav Kurma, Silpa Vadakkeeveetil Sreelatha, Manasi Patwardhan, and Shirish Karande 12:00–13:00

We consider a task based on CVPR 2018 challenge dataset on advertisement (Ad) understanding. The task involves detecting the viewer's interpretation of an Ad image captured as text. Recent results have shown that the embedded scene-text in the image holds a vital cue for this task. Motivated by this, we fine-tune the base BERT model for a sentence-pair classification task. Despite utilizing the scene-text as the only source of visual information, we could achieve a hit-or-miss accuracy of 84.95% on the challenge test data. To enable BERT to process other visual information, we append image captions to the scene-text. This achieves an accuracy of 89.69%, which is an improvement of 4.7%. This is the best reported result for this task.

Session 13A Semantics: Lexical-7

[CL] LESSLEX: Linking Multilingual Embeddings to SenSe Representations of Lexical Items [Website][PDF]
Davide Colla, Enrico Mensa, and Daniele P. Radicioni 12:00–13:00

We present LESSLEX, a novel multilingual lexical resource. Different from the vast majority of existing approaches, we ground our embeddings on a sense inventory made available from the BabelNet semantic network. In this setting, multilingual access is governed by the mapping of terms onto their underlying sense descriptions, such that all vectors co-exist in the same semantic space. As a result, for each term we have thus the 'blended' terminological vector along with those describing all senses associated to that term. LessLex has been tested on three tasks relevant to lexical semantics: conceptual similarity, contextual similarity, and semantic text similarity: we experimented over the principal data sets for such tasks in their multilingual and cross-lingual variants, improving on or closely approaching state-of-the-art results. We conclude by arguing that LessLex vectors may be relevant for practical applications and for research on conceptual and lexical access and competence.

Non-Linear Instance-Based Cross-Lingual Mapping for Non-Isomorphic Embedding Spaces [Website][PDF]
Goran Glavaš and Ivan Vulić 12:00–13:00

We present InstaMap, an instance-based method for learning projection-based cross-lingual word embeddings. Unlike prior work, it deviates from learning a single global linear projection. InstaMap is a non-parametric model that learns a non-linear projection by iteratively: (1) finding a globally optimal rotation of the source embedding space relying on the Kabsch algorithm, and then (2) moving each point along an instance-specific translation vector estimated from the translation vectors of the point's nearest neighbours in the training dictionary. We report performance gains with InstaMap over four representative state-of-the-art projection-based models on bilingual lexicon induction across a set of 28 diverse language pairs. We note prominent improvements, especially for more distant language pairs (i.e., languages with non-isomorphic monolingual spaces).

Session 13A Semantics: Sentence Level-8

Cross-Lingual Semantic Role Labeling with High-Quality Translated Training Corpus [Website][PDF]
Hao Fei, Meishan Zhang, and Donghong Ji 12:00–13:00

Many efforts of research are devoted to semantic role labeling (SRL) which is crucial for natural language understanding. Supervised approaches have achieved impressing performances when large-scale corpora are available for resource-rich languages such as English. While for the low-resource languages with no annotated SRL dataset, it is still challenging to obtain competitive performances. Cross-lingual SRL is one promising way to address the problem, which has achieved great advances with the help of model transferring and annotation projection. In this paper, we propose a novel alternative based on corpus translation, constructing high-quality training datasets for the target languages from the source gold-standard SRL annotations. Experimental results on Universal Proposition Bank show that the translation-based method is highly effective, and the automatic pseudo datasets can improve the target-language SRL performances significantly.

FastBERT: a Self-distilling BERT with Adaptive Inference Time [Website][PDF]
Weijie Liu, Peng Zhou, Zhiruo Wang, Zhe Zhao, Haotang Deng, and QI JU 12:00–13:00

Pre-trained language models like BERT have proven to be highly performant. However, they are often computationally expensive in many practical scenarios, for such heavy models can hardly be readily implemented with limited resources. To improve their efficiency with an assured model performance, we propose a novel speed-tunable FastBERT with adaptive inference time. The speed at inference can be flexibly adjusted under varying demands, while redundant calculation of samples is avoided. Moreover, this model adopts a unique self-distillation mechanism at fine-tuning, further enabling a greater computational efficacy with minimal loss in performance. Our model achieves promising results in twelve English and Chinese datasets. It is able to speed up by a wide range from 1 to 12 times than BERT if given different speedup thresholds to make a speed-performance tradeoff.

Good-Enough Compositional Data Augmentation [Website][PDF]
Jacob Andreas 12:00–13:00

We propose a simple data augmentation protocol aimed at providing a compositional inductive bias in conditional and unconditional sequence models. Under this protocol, synthetic training examples are constructed by taking real training examples and replacing (possibly discontinuous) fragments with other fragments that appear in at least one similar environment. The protocol is model-agnostic and useful for a variety of tasks. Applied to neural sequence-to-sequence models, it reduces error rate by as much as 87% on diagnostic tasks from the SCAN dataset and 16% on a semantic parsing task. Applied to n-gram language models, it reduces perplexity by roughly 1% on small corpora in several languages.

LogicalFactChecker: Leveraging Logical Operations for Fact Checking with Graph Module Network [Website][PDF]
Wanjun Zhong, Duyu Tang, Zhangyin Feng, Nan Duan, Ming Zhou, Ming Gong, Linjun Shou, Daxin Jiang, Jiahai Wang, and Jian Yin 12:00–13:00

Verifying the correctness of a textual statement requires not only semantic reasoning about the meaning of words, but also symbolic reasoning about logical operations like count, superlative, aggregation, etc. In this work, we propose LogicalFactChecker, a neural network approach capable of leveraging logical operations for fact checking. It achieves the state-of-the-art performance on TABFACT, a large-scale, benchmark dataset built for verifying a textual statement with semi-structured tables. This is achieved by a graph module network built upon the Transformer-based architecture. With a textual statement and a table as the input, LogicalFactChecker automatically derives a program (a.k.a. logical form) of the statement in a semantic parsing manner. A heterogeneous graph is then constructed to capture not only the structures of the table and the program, but also the connections between inputs with different modalities. Such a graph reveals the related contexts of each word in the statement, the table and the program. The graph is used to obtain graph-enhanced contextual representations of words in Transformer-based architecture. After that, a program-driven module network is further introduced to exploit the hierarchical structure of the program, where semantic compositionality is dynamically modeled along the program structure with a set of function-specific modules. Ablation experiments suggest that both the heterogeneous graph and the module network are important to obtain strong results.

Parsing into Variable-in-situ Logico-Semantic Graphs [Website][PDF]
Yufei Chen and Weiwei Sun 12:00–13:00

We propose variable-in-situ logico-semantic graphs to bridge the gap between semantic graph and logical form parsing. The new type of graph-based meaning representation allows us to include analysis for scope-related phenomena, such as quantification, negation and modality, in a way that is consistent with the state-of-the-art underspecification approach. Moreover, the well-formedness of such a graph is clear, since model-theoretic interpretation is available. We demonstrate the effectiveness of this new perspective by developing a new state-of-the-art semantic parser for English Resource Semantics. At the core of this parser is a novel neural graph rewriting system which combines the strengths of Hyperedge Replacement Grammar, a knowledge-intensive model, and Graph Neural Networks, a data-intensive model. Our parser achieves an accuracy of 92.39% in terms of elementary dependency match, which is a 2.88 point improvement over the best data-driven model in the literature. The output of our parser is highly coherent: at least 91% graphs are valid, in that they allow at least one sound scope-resolved logical form.

RAT-SQL: Relation-Aware Schema Encoding and Linking for Text-to-SQL Parsers [Website][PDF]
Bailin Wang, Richard Shin, Xiaodong Liu, Oleksandr Polozov, and Matthew Richardson 12:00–13:00

When translating natural language questions into SQL queries to answer questions from a database, contemporary semantic parsing models struggle to generalize to unseen database schemas. The generalization challenge lies in (a) encoding the database relations in an accessible way for the semantic parser, and (b) modeling alignment between database columns and their mentions in a given query. We present a unified framework, based on the relation-aware self-attention mechanism, to address schema encoding, schema linking, and feature representation within a text-to-SQL encoder. On the challenging Spider dataset this framework boosts the exact match accuracy to 57.2%, surpassing its best counterparts by 8.7% absolute improvement. Further augmented with BERT, it achieves the new state-of-the-art performance of 65.6% on the Spider leaderboard. In addition, we observe qualitative improvements in the model's understanding of schema linking and alignment. Our implementation will be open-sourced at https://github.com/Microsoft/rat-sql.

Semi-Supervised Semantic Dependency Parsing Using CRF Autoencoders [Website][PDF]
Zixia Jia, Youmi Ma, Jiong Cai, and Kewei Tu 12:00–13:00

Semantic dependency parsing, which aims to find rich bi-lexical relationships, allows words to have multiple dependency heads, resulting in graph-structured representations. We propose an approach to semi-supervised learning of semantic dependency parsers based on the CRF autoencoder framework. Our encoder is a discriminative neural semantic dependency parser that predicts the latent parse graph of the input sentence. Our decoder is a generative neural model that reconstructs the input sentence conditioned on the latent parse graph. Our model is arc-factored and therefore parsing and learning are both tractable. Experiments show our model achieves significant and consistent improvement over the supervised baseline.

Sentence Meta-Embeddings for Unsupervised Semantic Textual Similarity [Website][PDF]
Nina Poerner, Ulli Waltinger, and Hinrich Schütze 12:00–13:00

We address the task of unsupervised Semantic Textual Similarity (STS) by ensembling diverse pre-trained sentence encoders into sentence meta-embeddings. We apply, extend and evaluate different meta-embedding methods from the word embedding literature at the sentence level, including dimensionality reduction (Yin and Schütze, 2016), generalized Canonical Correlation Analysis (Rastogi et al., 2015) and cross-view auto-encoders (Bollegala and Bao, 2018). Our sentence meta-embeddings set a new unsupervised State of The Art (SoTA) on the STS Benchmark and on the STS12-STS16 datasets, with gains of between 3.7% and 6.4% Pearson's r over single-source systems.

Transition-based Semantic Dependency Parsing with Pointer Networks [Website][PDF]
Daniel Fernández-González and Carlos Gómez-Rodríguez 12:00–13:00

Transition-based parsers implemented with Pointer Networks have become the new state of the art in dependency parsing, excelling in producing labelled syntactic trees and outperforming graph-based models in this task. In order to further test the capabilities of these powerful neural networks on a harder NLP problem, we propose a transition system that, thanks to Pointer Networks, can straightforwardly produce labelled directed acyclic graphs and perform semantic dependency parsing. In addition, we enhance our approach with deep contextualized word embeddings extracted from BERT. The resulting system not only outperforms all existing transition-based models, but also matches the best fully-supervised accuracy to date on the SemEval 2015 Task 18 datasets among previous state-of-the-art graph-based parsers.

tBERT: Topic Models and BERT Joining Forces for Semantic Similarity Detection [Website][PDF]
Nicole Peinelt, Dong Nguyen, and Maria Liakata 12:00–13:00

Semantic similarity detection is a fundamental task in natural language understanding. Adding topic information has been useful for previous feature-engineered semantic similarity models as well as neural models for other tasks. There is currently no standard way of combining topics with pretrained contextual representations such as BERT. We propose a novel topic-informed BERT-based architecture for pairwise semantic similarity detection and show that our model improves performance over strong neural baselines across a variety of English language datasets. We find that the addition of topics to BERT helps particularly with resolving domain-specific cases.

Session 13A Semantics: Textual Inference and Other Areas of Semantics-5

Curriculum Learning for Natural Language Understanding [Website][PDF]
Benfeng Xu, Licheng Zhang, Zhendong Mao, Quan Wang, Hongtao Xie, and Yongdong Zhang 12:00–13:00

With the great success of pre-trained language models, the pretrain-finetune paradigm now becomes the undoubtedly dominant solution for natural language understanding (NLU) tasks. At the fine-tune stage, target task data is usually introduced in a completely random order and treated equally. However, examples in NLU tasks can vary greatly in difficulty, and similar to human learning procedure, language models can benefit from an easy-to-difficult curriculum. Based on this idea, we propose our Curriculum Learning approach. By reviewing the trainset in a crossed way, we are able to distinguish easy examples from difficult ones, and arrange a curriculum for language models. Without any manual model architecture design or use of external data, our Curriculum Learning approach obtains significant and universal performance improvements on a wide range of NLU tasks.

Evidence-Aware Inferential Text Generation with Vector Quantised Variational AutoEncoder [Website][PDF]
Daya Guo, Duyu Tang, Nan Duan, Jian Yin, Daxin Jiang, and Ming Zhou 12:00–13:00

Generating inferential texts about an event in different perspectives requires reasoning over different contexts that the event occurs. Existing works usually ignore the context that is not explicitly provided, resulting in a context-independent semantic representation that struggles to support the generation. To address this, we propose an approach that automatically finds evidence for an event from a large text corpus, and leverages the evidence to guide the generation of inferential texts. Our approach works in an encoderdecoder manner and is equipped with Vector Quantised-Variational Autoencoder, where the encoder outputs representations from a distribution over discrete variables. Such discrete representations enable automatically selecting relevant evidence, which not only facilitates evidence-aware generation, but also provides a natural way to uncover rationales behind the generation. Our approach provides state-of-the-art performance on both Event2mind and Atomic datasets. More importantly, we find that with discrete representations, our model selectively uses evidence to generate different inferential texts.

Fine-grained Fact Verification with Kernel Graph Attention Network [Website][PDF]
Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu 12:00–13:00

Fact Verification requires fine-grained natural language inference capability that finds subtle clues to identify the syntactical and semantically correct but not well-supported claims. This paper presents Kernel Graph Attention Network (KGAT), which conducts more fine-grained fact verification with kernel-based attentions. Given a claim and a set of potential evidence sentences that form an evidence graph, KGAT introduces node kernels, which better measure the importance of the evidence node, and edge kernels, which conduct fine-grained evidence propagation in the graph, into Graph Attention Networks for more accurate fact verification. KGAT achieves a 70.38% FEVER score and significantly outperforms existing fact verification models on FEVER, a large-scale benchmark for fact verification. Our analyses illustrate that, compared to dot-product attentions, the kernel-based attention concentrates more on relevant evidence sentences and meaningful clues in the evidence graph, which is the main source of KGAT's effectiveness. All source codes of this work are available at https://github.com/thunlp/KernelGAT.

NeuInfer: Knowledge Inference on N-ary Facts [Website][PDF]
Saiping Guan, Xiaolong Jin, Jiafeng Guo, Yuanzhuo Wang, and Xueqi Cheng 12:00–13:00

Knowledge inference on knowledge graph has attracted extensive attention, which aims to find out connotative valid facts in knowledge graph and is very helpful for improving the performance of many downstream applications. However, researchers have mainly poured attention to knowledge inference on binary facts. The studies on n-ary facts are relatively scarcer, although they are also ubiquitous in the real world. Therefore, this paper addresses knowledge inference on n-ary facts. We represent each n-ary fact as a primary triple coupled with a set of its auxiliary descriptive attribute-value pair(s). We further propose a neural network model, NeuInfer, for knowledge inference on n-ary facts. Besides handling the common task to infer an unknown element in a whole fact, NeuInfer can cope with a new type of task, flexible knowledge inference. It aims to infer an unknown element in a partial fact consisting of the primary triple coupled with any number of its auxiliary description(s). Experimental results demonstrate the remarkable superiority of NeuInfer.

Neural Graph Matching Networks for Chinese Short Text Matching [Website][PDF]
Lu Chen, Yanbin Zhao, Boer Lyu, Lesheng Jin, Zhi Chen, Su Zhu, and Kai Yu 12:00–13:00

Chinese short text matching usually employs word sequences rather than character sequences to get better performance. However, Chinese word segmentation can be erroneous, ambiguous or inconsistent, which consequently hurts the final matching performance. To address this problem, we propose neural graph matching networks, a novel sentence matching framework capable of dealing with multi-granular input information. Instead of a character sequence or a single word sequence, paired word lattices formed from multiple word segmentation hypotheses are used as input and the model learns a graph representation according to an attentive graph matching mechanism. Experiments on two Chinese datasets show that our models outperform the state-of-the-art short text matching models.

Premise Selection in Natural Language Mathematical Texts [Website][PDF]
Deborah Ferreira and André Freitas 12:00–13:00

The discovery of supporting evidence for addressing complex mathematical problems is a semantically challenging task, which is still unexplored in the field of natural language processing for mathematical text. The natural language premise selection task consists in using conjectures written in both natural language and mathematical formulae to recommend premises that most likely will be useful to prove a particular statement. We propose an approach to solve

this task as a link prediction problem, using Deep Convolutional Graph Neural Networks. This paper also analyses how different baselines perform in this task and shows that a graph structure can provide higher F1-score, especially when considering multi-hop premise selection.

Reasoning Over Semantic-Level Graph for Fact Checking [Website][PDF]
Wanjun Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin
12:00–13:00

Fact checking is a challenging task because verifying the truthfulness of a claim requires reasoning about multiple retrievable evidence. In this work, we present a method suitable for reasoning about the semantic-level structure of evidence. Unlike most previous works, which typically represent evidence sentences with either string concatenation or fusing the features of isolated evidence sentences, our approach operates on rich semantic structures of evidence obtained by semantic role labeling. We propose two mechanisms to exploit the structure of evidence while leveraging the advances of pre-trained models like BERT, GPT or XLNet. Specifically, using XLNet as the backbone, we first utilize the graph structure to re-define the relative distances of words, with the intuition that semantically related words should have short distances. Then, we adopt graph convolutional network and graph attention network to propagate and aggregate information from neighboring nodes on the graph. We evaluate our system on FEVER, a benchmark dataset for fact checking, and find that rich structural information is helpful and both our graph-based mechanisms improve the accuracy. Our model is the state-of-the-art system in terms of both official evaluation metrics, namely claim verification accuracy and FEVER score.

Temporal Common Sense Acquisition with Minimal Supervision [Website][PDF]
Ben Zhou, Qiang Ning, Daniel Khashabi, and Dan Roth 12:00–13:00

Temporal common sense (e.g., duration and frequency of events) is crucial for understanding natural language. However, its acquisition is challenging, partly because such information is often not expressed explicitly in text, and human annotation on such concepts is costly. This work proposes a novel sequence modeling approach that exploits explicit and implicit mentions of temporal common sense, extracted from a large corpus, to build TacoLM, a temporal common sense language model. Our method is shown to give quality predictions of various dimensions of temporal common sense (on UDST and a newly collected dataset from RealNews). It also produces representations of events for relevant tasks such as duration comparison, parent-child relations, event coreference and temporal QA (on TimeBank, HiEVE and MCTACO) that are better than using the standard BERT. Thus, it will be an important component of temporal NLP.

The Sensitivity of Language Models and Humans to Winograd Schema Perturbations [Website][PDF]
Mostafa Abdou, Vinit Ravishankar, Maria Barrett, Yonatan Belinkov, Desmond Elliott, and Anders Søgaard 12:00–13:00

Large-scale pretrained language models are the major driving force behind recent improvements in perfromance on the Winograd Schema Challenge, a widely employed test of commonsense reasoning ability. We show, however, with a new diagnostic dataset, that these models are sensitive to linguistic perturbations of the Winograd examples that minimally affect human understanding. Our results highlight interesting differences between humans and language models: language models are more sensitive to number or gender alternations and synonym replacements than humans, and humans are more stable and consistent in their predictions, maintain a much higher absolute performance, and perform better on non-associative instances than associative ones.

Session 13A: Sentiment Analysis, Stylistic Analysis, and Argument Mining-11

Conditional Augmentation for Aspect Term Extraction via Masked Sequence-to-Sequence Generation [Website][PDF]
Kun Li, Chengbo Chen, Xiaojun Quan, Qing Ling, and Yan Song 12:00–13:00

Aspect term extraction aims to extract aspect terms from review texts as opinion targets for sentiment analysis. One of the big challenges with this task is the lack of sufficient annotated data. While data augmentation is potentially an effective technique to address the above issue, it is uncontrollable as it may change aspect words and aspect labels unexpectedly. In this paper, we formulate the data augmentation as a conditional generation task: generating a new sentence while preserving the original opinion targets and labels. We propose a masked sequence-to-sequence method for conditional augmentation of aspect term extraction. Unlike existing augmentation approaches, ours is controllable and allows to generate more diversified sentences. Experimental results confirm that our method alleviates the data scarcity problem significantly. It also effectively boosts the performances of several current models for aspect term extraction.

Exploiting Personal Characteristics of Debaters for Predicting Persuasiveness [Website][PDF]
Khalid Al Khatib, Michael Völske, Shahbaz Syed, Nikolay Kolyada, and Benno Stein 12:00–13:00

Predicting the persuasiveness of arguments has applications as diverse as writing assistance, essay scoring, and advertising. While clearly relevant to the task, the personal characteristics of an argument's source and audience have not yet been fully exploited toward automated persuasiveness prediction. In this paper, we model debaters' prior beliefs, interests, and personality traits based on their previous activity, without dependence on explicit user profiles or questionnaires. Using a dataset of over 60,000 argumentative discussions, comprising more than three million individual posts collected from the subreddit r/ChangeMyView, we demonstrate that our modeling of debater's characteristics enhances the prediction of argument persuasiveness as well as of debaters' resistance to persuasion.

Out of the Echo Chamber: Detecting Countering Debate Speeches [Website][PDF]
Matan Orbach, Yonatan Bilu, Assaf Toledo, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim 12:00–13:00

An educated and informed consumption of media content has become a challenge in modern times. With the shift from traditional news outlets to social media and similar venues, a major concern is that readers are becoming encapsulated in "echo chambers" and may fall prey to fake news and disinformation, lacking easy access to dissenting views. We suggest a novel task aiming to alleviate some of these concerns – that of detecting articles that most effectively counter the arguments – and not just the stance – made in a given text. We study this problem in the context of debate speeches. Given such a speech, we aim to identify, from among a set of speeches on the same topic and with an opposing stance, the ones that directly counter it. We provide a large dataset of 3,685 such speeches (in English), annotated for this relation, which hopefully would be of general interest to the NLP community. We explore several algorithms addressing this task, and while some are successful, all fall short of expert human performance, suggesting room for further research. All data collected during this work is freely available for research.

Session 13A: Student Research Workshop

Pre-training via Leveraging Assisting Languages for Neural Machine Translation [Website][PDF]
Haiyue Song, Raj Dabre, Zhuoyuan Mao, Fei Cheng, Sadao Kurohashi, and Eiichiro Sumita 12:00–13:00

Sequence-to-sequence (S2S) pre-training using large monolingual data is known to improve performance for various S2S NLP tasks. However, large monolingual corpora might not always be available for the languages of interest (LOI). Thus, we propose to exploit monolingual corpora of other languages to complement the scarcity of monolingual corpora for the LOI. We utilize script mapping (Chinese to Japanese) to increase the similarity (number of cognates) between the monolingual corpora of helping languages and LOI. An empirical case study of low-resource Japanese-English neural machine translation (NMT) reveals that leveraging large Chinese and French monolingual corpora can help overcome the shortage of Japanese and English monolingual corpora, respectively, for S2S pre-training. Using only Chinese and French monolingual corpora, we were able to improve Japanese-English translation quality by up to 8.5 BLEU in low-resource scenarios.

A Simple and Effective Dependency Parser for Telugu [Website][PDF]
Sneha Nallani, Manish Shrivastava, and Dipti Sharma 12:00–13:00

We present a simple and effective dependency parser for Telugu, a morphologically rich, free word order language. We propose to replace the rich linguistic feature templates used in the past approaches with a minimal feature function using contextual vector representations. We train a BERT model on the Telugu Wikipedia data and use vector representations from this model to train the parser. Each sentence token is associated with a vector representing the token in the context of that sentence and the feature vectors are constructed by concatenating two token representations from the stack and one from the buffer. We put the feature representations through a feedforward network and train with a greedy transition based approach. The resulting parser has a very simple architecture with minimal feature engineering and achieves state-of-the-art results for Telugu.

Cross-Lingual Disaster-related Multi-label Tweet Classification with Manifold Mixup [Website][PDF]
Jishnu Ray Chowdhury, Cornelia Caragea, and Doina Caragea 12:00–13:00

Distinguishing informative and actionable messages from a social media platform like Twitter is critical for facilitating disaster management. For this purpose, we compile a multilingual dataset of over 130K samples for multi-label classification of disaster-related tweets. We present a masking-based loss function for partially labelled samples and demonstrate the effectiveness of Manifold Mixup in the text domain. Our main model is based on Multilingual BERT, which we further improve with Manifold Mixup. We show that our model generalizes to unseen disasters in the test set. Furthermore, we analyze the capability of our model for zero-shot generalization to new languages. Our code, dataset, and other resources are available on Github.

Demo Session 3B

Time: *12:45–13:30*

Clinical-Coder: Assigning Interpretable ICD-10 Codes to Chinese Clinical Notes [Website][PDF]
Pengfei Cao, Chenwei Yan, Xiangling Fu, Yubo Chen, Kang Liu, Jun Zhao, Shengping Liu, and Weifeng Chong

In this paper, we introduce Clinical-Coder, an online system aiming to assign ICD codes to Chinese clinical notes. ICD coding has been a research hotspot of clinical medicine, but the interpretability of prediction hinders its practical application. We exploit a Dilated Convolutional Attention network with N-gram Matching mechanism (DCANM) to capture semantic features for non-continuous words and continuous n-gram words, concentrating on explaining the reason why each ICD code to be predicted. The experiments demonstrate that our approach is effective and that our system is able to provide supporting information in clinical decision making.

Prta: A System to Support the Analysis of Propaganda Techniques in the News [Website][PDF]
Giovanni Da San Martino, Shaden Shaar, Yifan Zhang, Seunghak Yu, Alberto Barrón-Cedeño, and Preslav Nakov

Recent events, such as the 2016 US Presidential Campaign, Brexit and the COVID-19 "infodemic", have brought into the spotlight the dangers of online disinformation. There has been a lot of research focusing on fact-checking and disinformation detection. However, little attention has been paid to the specific rhetorical and psychological techniques used to convey propaganda messages. Revealing the use of such techniques can help promote media literacy and critical thinking, and eventually contribute to limiting the impact of "fake news" and disinformation campaigns. Prta (Propaganda Persuasion Techniques Analyzer) allows users to explore the articles crawled on a regular basis by highlighting the spans in which propaganda techniques occur and to compare them on the basis of their use of propaganda techniques. The system further reports statistics about the use of such techniques, overall and over time, or according to filtering criteria specified by the user based on time interval, keywords, and/or political orientation of the media. Moreover, it allows users to analyze any text or URL through a dedicated interface or via an API. The system is available online: https://www.tanbih.org/prta.

NSTM: Real-Time Query-Driven News Overview Composition at Bloomberg [Website][PDF]
Joshua Bambrick, Minjie Xu, Andy Almonte, Igor Malioutov, Guim Perarnau, Vittorio Selo, and Iat Chong Chan

Millions of news articles from hundreds of thousands of sources around the globe appear in news aggregators every day. Consuming such a volume of news presents an almost insurmountable challenge. For example, a reader searching on Bloomberg's system for news about the U.K. would find 10,000 articles on a typical day. Apple Inc., the world's most journalistically covered company, garners around 1,800 news articles a day. We realized that a new kind of summarization engine was needed, one that would condense large volumes of news into short, easy to absorb points. The system would filter out noise and duplicates to identify and summarize key news about companies, countries or markets. When given a user query, Bloomberg's solution, Key News Themes (or NSTM), leverages state-of-the-art semantic clustering techniques and novel summarization methods to produce comprehensive, yet concise, digests to dramatically simplify the news consumption process. NSTM is available to hundreds of thousands of readers around the world and serves thousands of requests daily with sub-second latency. At ACL 2020, we will present a demo of NSTM.

Session 13B Overview – Wednesday, July 8, 2020 13:00–14:00

Track A *Dialogue and Interactive Systems-16* Abstracts	A Contextual Hierarchical Attention Network with Adaptive Objective for Dialogue State Tracking *Shan, Li, Zhang, Meng, Feng, Niu, and Zhou* [Website][PDF]	Diversifying Dialogue Generation with Non-Conversational Text *Su, Shen, Zhao, Xiao, Hu, Niu, and Zhou* [Website][PDF]	Dynamic Fusion Network for Multi-Domain End-to-end Task-Oriented Dialog *Qin, Xu, Che, Zhang, and Liu* [Website][PDF]	KdConv: A Chinese Multi-domain Dialogue Dataset Towards Multi-turn Knowledge-driven Conversation *Zhou, Zheng, Huang, Huang, and Zhu* [Website][PDF]	Learning to Customize Model Structures for Few-shot Dialogue Generation Tasks *SONG, Liu, Bi, Yan, and Zhang* [Website][PDF]
	Meta-Reinforced Multi-Domain State Generator for Dialogue Systems *Huang, Feng, Hu, Wu, Du, and Ma* [Website][PDF]	Modeling Long Context for Task-Oriented Dialogue State Generation *Quan and Xiong* [Website][PDF]	Multi-Domain Dialogue Acts and Response Co-Generation *Wang, Tian, Wang, Quan, and Yu* [Website][PDF]	Speaker Sensitive Response Evaluation Model *Bak and Oh* [Website][PDF]	
Track B *Discourse and Pragmatics-8* Abstracts	DRTS Parsing with Structure-Aware Encoding and Decoding *Fu, Zhang, Liu, and Zhang* [Website][PDF]				
Track C *Information Extraction-8* Abstracts	Bipartite Flat-Graph Network for Nested Named Entity Recognition *Luo and Zhao* [Website][PDF]	FLAT: Chinese NER Using Flat-Lattice Transformer *Li, Yan, Qiu, and Huang* [Website][PDF]	Temporally-Informed Analysis of Named Entity Recognition *Rijhwani and Preotiuc-Pietro* [Website][PDF]	Towards Open Domain Event Trigger Identification using Adversarial Domain Adaptation *Naik and Rose* [Website][PDF]	
Track D *Language Grounding to Vision, Robotics and Beyond-7* Abstracts	Aligned Dual Channel Graph Convolutional Network for Visual Question Answering *Huang, Wei, Cai, Zheng, Chen, Leung, and Li* [Website][PDF]	CompGuessWhat?!: A Multi-task Evaluation Framework for Grounded Language Learning *Suglia, Konstas, Vanzo, Bastianelli, Elliott, Frank, and Lemon* [Website][PDF]	Cross-Modality Relevance for Reasoning on Language and Vision *Zheng, Guo, and Kordjamshidi* [Website][PDF]	Learning Web-based Procedures by Reasoning over Explanations and Demonstrations in Context *Srivastava, Polozov, Jojic, and Meek* [Website][PDF]	Multi-agent Communication meets Natural Language: Synergies between Functional and Structural Language Learning *Lazaridou, Potapenko, and Tieleman* [Website][PDF]
	Multimodal Neural Graph Memory Networks for Visual Question Answering *Khademi* [Website][PDF]				
Track E *Machine Translation-15* Abstracts	HAT: Hardware-Aware Transformers for Efficient Natural Language Processing *Wang, Wu, Liu, Cai, Zhu, Gan, and Han* [Website][PDF]	Hard-Coded Gaussian Attention for Neural Machine Translation *You, Sun, and Iyyer* [Website][PDF]	In Neural Machine Translation, What Does Transfer Learning Transfer? *Aji, Bogoychev, Heafield, and Sennrich* [Website][PDF]	Learning a Multi-Domain Curriculum for Neural Machine Translation *Wang, Tian, Ngiam, Yang, Caswell, and Parekh* [Website][PDF]	Reducing Gender Bias in Neural Machine Translation as a Domain Adaptation Problem *Saunders and Byrne* [Website][PDF]

	Translationese as a Language in "Multilingual" NMT *Riley, Caswell, Freitag, and Grangier* [Website][PDF]	Uncertainty-Aware Curriculum Learning for Neural Machine Translation *Zhou, Yang, Wong, Wan, and Chao* [Website][PDF]	Unsupervised Domain Clusters in Pretrained Language Models *Aharoni and Goldberg* [Website][PDF]	Using Context in Neural Machine Translation Training Objectives *Saunders, Stahlberg, and Byrne* [Website][PDF]	Variational Neural Machine Translation with Normalizing Flows *Setiawan, Sperber, Nallasamy, and Paulik* [Website][PDF]
Track F *Phonology, Morphology and Word Segmentation-5* Abstracts	2kenize: Tying Subword Sequences for Chinese Script Conversion *Pranav A and Augenstein* [Website][PDF]	[TACL] Phonotactic Complexity and its Trade-offs *Pimentel, Roark, and Cotterell* [Website][PDF]	Predicting the Growth of Morphological Families from Social and Linguistic Factors *Hofmann, Pierrehumbert, and Schütze* [Website][PDF]	Semi-supervised Contextual Historical Text Normalization *Makarov and Clematide* [Website][PDF]	The Paradigm Discovery Problem *Erdmann, Elsner, Wu, Cotterell, and Habash* [Website][PDF]
	Supervised Grapheme-to-Phoneme Conversion of Orthographic Schwas in Hindi and Punjabi *Arora, Gessler, and Schneider* [Website][PDF]				
Track G *Question Answering-11* Abstracts	ClarQ: A large-scale and diverse dataset for Clarification Question Generation *Kumar and Black* [Website][PDF]	DoQA - Accessing Domain-Specific FAQs via Conversational QA *Campos, Otegi, Soroa, Deriu, Cieliebak, and Agirre* [Website][PDF]	Document Modeling with Graph Attention Networks for Multi-grained Machine Reading Comprehension *Zheng, Wen, Liang, Duan, Che, Jiang, Zhou, and Liu* [Website][PDF]	Low-Resource Generation of Multi-hop Reasoning Questions *Yu, Liu, Qiu, Su, Wang, Quan, and Yin* [Website][PDF]	MLQA: Evaluating Cross-lingual Extractive Question Answering *Lewis, Oguz, Rinott, Riedel, and Schwenk* [Website][PDF]
	Multi-source Meta Transfer for Low Resource Multiple-Choice Question Answering *Yan, Zhang, Jin, and Zhou* [Website][PDF]	R4C: A Benchmark for Evaluating RC Systems to Get the Right Answer for the Right Reason *Inoue, Stenetorp, and Inui* [Website][PDF]	Recurrent Chunking Mechanisms for Long-Text Machine Reading Comprehension *Gong, Shen, Yu, Chen, and Yu* [Website][PDF]		
Track H *Student Research Workshop* Abstracts	AraDIC: Arabic Document Classification Using Image-Based Character Embeddings and Class-Balanced Loss *Daif, Kitada, and Iyatomi* [Website][PDF]	Understanding Points of Correspondence between Sentences for Abstractive Summarization *Lebanoff, Muchovej, Dernoncourt, Kim, Wang, Chang, and Liu* [Website][PDF]	Noise-Based Augmentation Techniques for Emotion Datasets: What do we Recommend? *Jaiswal and Provost* [Website]	Logical Inferences with Comparatives and Generalized Quantifiers *Haruta, Mineshima, and Bekki* [Website][PDF]	
Track I *Theme-5* Abstracts	A Call for More Rigor in Unsupervised Cross-lingual Learning *Artetxe, Ruder, Yogatama, Labaka, and Agirre* [Website][PDF]	A Tale of a Probe and a Parser *Hall Maudslay, Valvoda, Pimentel, Williams, and Cotterell* [Website][PDF]	Are we Estimating or Guesstimating Translation Quality? *Sun, Guzmán, and Specia* [Website][PDF]	Automated Evaluation of Writing — 50 Years and Counting *Beigman Klebanov and Madnani* [Website][PDF]	From SPMRL to NMRL: What Did We Learn (and Unlearn) in a Decade of Parsing Morphologically-Rich Languages (MRLs)? *Tsarfaty, Bareket, Klein, and Seker* [Website][PDF]

Language (Re)modelling: Towards Embodied Language Understanding *Tamari, Shani, Hope, Petruck, Abend, and Shahaf* [Website][PDF]	Negated and Misprimed Probes for Pretrained Language Models: Birds Can Talk, But Cannot Fly *Kassner and Schütze* [Website][PDF]	On Forgetting to Cite Older Papers: An Analysis of the ACL Anthology *Bollmann and Elliott* [Website][PDF]	Returning the N to NLP: Towards Contextually Personalized Classification Models *Flek* [Website][PDF]	Speech Translation and the End-to-End Promise: Taking Stock of Where We Are *Sperber and Paulik* [Website][PDF]
To Test Machine Comprehension, Start by Defining Comprehension *Dunietz, Burnham, Bharadwaj, Rambow, Chu-Carroll, and Ferrucci* [Website][PDF]	What Question Answering can Learn from Trivia Nerds *Boyd-Graber and Börschinger* [Website][PDF]	What are the Goals of Distributional Semantics? *Emerson* [Website][PDF]	Gender Gap in Natural Language Processing Research: Disparities in Authorship and Citations *Mohammad* [Website][PDF]	

Session 13B Details

Session 13B: Dialogue and Interactive Systems-16

A Contextual Hierarchical Attention Network with Adaptive Objective for Dialogue State Tracking [Website][PDF]
Yong Shan, Zekang Li, Jinchao Zhang, Fandong Meng, Yang Feng, Cheng Niu, and Jie Zhou 13:00–14:00

Recent studies in dialogue state tracking (DST) leverage historical information to determine states which are generally represented as slot-value pairs. However, most of them have limitations to efficiently exploit relevant context due to the lack of a powerful mechanism for modeling interactions between the slot and the dialogue history. Besides, existing methods usually ignore the slot imbalance problem and treat all slots indiscriminately, which limits the learning of hard slots and eventually hurts overall performance. In this paper, we propose to enhance the DST through employing a contextual hierarchical attention network to not only discern relevant information at both word level and turn level but also learn contextual representations. We further propose an adaptive objective to alleviate the slot imbalance problem by dynamically adjust weights of different slots during training. Experimental results show that our approach reaches 52.68% and 58.55% joint accuracy on MultiWOZ 2.0 and MultiWOZ 2.1 datasets respectively and achieves new state-of-the-art performance with considerable improvements (+1.24% and +5.98%).

Diversifying Dialogue Generation with Non-Conversational Text [Website][PDF]
Hui Su, Xiaoyu Shen, Sanqiang Zhao, Zhou Xiao, Pengwei Hu, randy zhong randy, Cheng Niu, and Jie Zhou 13:00–14:00

Neural network-based sequence-to-sequence (seq2seq) models strongly suffer from the low-diversity problem when it comes to open-domain dialogue generation. As bland and generic utterances usually dominate the frequency distribution in our daily chitchat, avoiding them to generate more interesting responses requires complex data filtering, sampling techniques or modifying the training objective. In this paper, we propose a new perspective to diversify dialogue generation by leveraging *non-conversational* text. Compared with bilateral conversations, non-conversational text are easier to obtain, more diverse and cover a much broader range of topics. We collect a large-scale non-conversational corpus from multi sources including forum comments, idioms and book snippets. We further present a training paradigm to effectively incorporate these text via iterative back translation. The resulting model is tested on two conversational datasets from different domains and is shown to produce significantly more diverse responses without sacrificing the relevance with context.

Dynamic Fusion Network for Multi-Domain End-to-end Task-Oriented Dialog [Website][PDF]
Libo Qin, Xiao Xu, Wanxiang Che, Yue Zhang, and Ting Liu 13:00–14:00

Recent studies have shown remarkable success in end-to-end task-oriented dialog system. However, most neural models rely on large training data, which are only available for a certain number of task domains, such as navigation and scheduling. This makes it difficult to scalable for a new domain with limited labeled data. However, there has been relatively little research on how to effectively use data from all domains to improve the performance of each domain and also unseen domains. To this end, we investigate methods that can make explicit use of domain knowledge and introduce a shared-private network to learn shared and specific knowledge. In addition, we propose a novel Dynamic Fusion Network (DF-Net) which automatically exploit the relevance between the target domain and each domain. Results show that our models outperforms existing methods on multi-domain dialogue, giving the state-of-the-art in the literature. Besides, with little training data, we show its transferability by outperforming prior best model by 13.9% on average.

KdConv: A Chinese Multi-domain Dialogue Dataset Towards Multi-turn Knowledge-driven Conversation [Website][PDF]
Hao Zhou, Chujie Zheng, Kaili Huang, Minlie Huang, and Xiaoyan Zhu 13:00–14:00

The research of knowledge-driven conversational systems is largely limited due to the lack of dialog data which consists of multi-turn conversations on multiple topics and with knowledge annotations. In this paper, we propose a Chinese multi-domain knowledge-driven conversation dataset, KdConv, which grounds the topics in multi-turn conversations to knowledge graphs. Our corpus contains 4.5K conversations from three domains (film, music, and travel), and 86K utterances with an average turn number of 19.0. These conversations contain in-depth discussions on related topics and natural transition between multiple topics. To facilitate the following research on this corpus, we provide several benchmark models. Comparative results show that the models can be enhanced by introducing background knowledge, yet there is still a large space for leveraging knowledge to model multi-turn conversations for further research. Results also show that there are obvious performance differences between different domains, indicating that it is worth further explore transfer learning and domain adaptation. The corpus and benchmark models are publicly available.

Learning to Customize Model Structures for Few-shot Dialogue Generation Tasks [Website][PDF]
YIPING SONG, Zequn Liu, Wei Bi, Rui Yan, and Ming Zhang 13:00–14:00

Training the generative models with minimal corpus is one of the critical challenges for building open-domain dialogue systems. Existing methods tend to use the meta-learning framework which pre-trains the parameters on all non-target tasks then fine-tunes on the target task. However, fine-tuning distinguishes tasks from the parameter perspective but ignores the model-structure perspective, resulting in similar dialogue models for different tasks. In this

paper, we propose an algorithm that can customize a unique dialogue model for each task in the few-shot setting. In our approach, each dialogue model consists of a shared module, a gating module, and a private module. The first two modules are shared among all the tasks, while the third one will differentiate into different network structures to better capture the characteristics of the corresponding task. The extensive experiments on two datasets show that our method outperforms all the baselines in terms of task consistency, response quality, and diversity.

Meta-Reinforced Multi-Domain State Generator for Dialogue Systems [Website][PDF]
Yi Huang, Junlan Feng, Min Hu, Xiaoting Wu, Xiaoyu Du, and Shuo Ma 13:00–14:00

A Dialogue State Tracker (DST) is a core component of a modular task-oriented dialogue system. Tremendous progress has been made in recent years. However, the major challenges remain. The state-of-the-art accuracy for DST is below 50% for a multi-domain dialogue task. A learnable DST for any new domain requires a large amount of labeled in-domain data and training from scratch. In this paper, we propose a Meta-Reinforced Multi-Domain State Generator (MERET). Our first contribution is to improve the DST accuracy. We enhance a neural model based DST generator with a reward manager, which is built on policy gradient reinforcement learning (RL) to fine-tune the generator. With this change, we are able to improve the joint accuracy of DST from 48.79% to 50.91% on the MultiWOZ corpus. Second, we explore to train a DST meta-learning model with a few domains as source domains and a new domain as target domain. We apply the model-agnostic meta-learning algorithm (MAML) to DST and the obtained meta-learning model is used for new domain adaptation. Our experimental results show this solution is able to outperform the traditional training approach with extremely less training data in target domain.

Modeling Long Context for Task-Oriented Dialogue State Generation [Website][PDF]
Jun Quan and Deyi Xiong 13:00–14:00

Based on the recently proposed transferable dialogue state generator (TRADE) that predicts dialogue states from utterance-concatenated dialogue context, we propose a multi-task learning model with a simple yet effective utterance tagging technique and a bidirectional language model as an auxiliary task for task-oriented dialogue state generation. By enabling the model to learn a better representation of the long dialogue context, our approaches attempt to solve the problem that the performance of the baseline significantly drops when the input dialogue context sequence is long. In our experiments, our proposed model achieves a 7.03% relative improvement over the baseline, establishing a new state-of-the-art joint goal accuracy of 52.04% on the MultiWOZ 2.0 dataset.

Multi-Domain Dialogue Acts and Response Co-Generation [Website][PDF]
Kai Wang, Junfeng Tian, Rui Wang, Xiaojun Quan, and Jianxing Yu 13:00–14:00

Generating fluent and informative responses is of critical importance for task-oriented dialogue systems. Existing pipeline approaches generally predict multiple dialogue acts first and use them to assist response generation. There are at least two shortcomings with such approaches. First, the inherent structures of multi-domain dialogue acts are neglected. Second, the semantic associations between acts and responses are not taken into account for response generation. To address these issues, we propose a neural co-generation model that generates dialogue acts and responses concurrently. Unlike those pipeline approaches, our act generation module preserves the semantic structures of multi-domain dialogue acts and our response generation module dynamically attends to different acts as needed. We train the two modules jointly using an uncertainty loss to adjust their task weights adaptively. Extensive experiments are conducted on the large-scale MultiWOZ dataset and the results show that our model achieves very favorable improvement over several state-of-the-art models in both automatic and human evaluations.

Speaker Sensitive Response Evaluation Model [Website][PDF]
JinYeong Bak and Alice Oh 13:00–14:00

Automatic evaluation of open-domain dialogue response generation is very challenging because there are many appropriate responses for a given context. Existing evaluation models merely compare the generated response with the ground truth response and rate many of the appropriate responses as inappropriate if they deviate from the ground truth. One approach to resolve this problem is to consider the similarity of the generated response with the conversational context. In this paper, we propose an automatic evaluation model based on that idea and learn the model parameters from an unlabeled conversation corpus. Our approach considers the speakers in defining the different levels of similar context. We use a Twitter conversation corpus that contains many speakers and conversations to test our evaluation model. Experiments show that our model outperforms the other existing evaluation metrics in terms of high correlation with human annotation scores. We also show that our model trained on Twitter can be applied to movie dialogues without any additional training. We provide our code and the learned parameters so that they can be used for automatic evaluation of dialogue response generation models.

Session 13B: Discourse and Pragmatics-8

DRTS Parsing with Structure-Aware Encoding and Decoding [Website][PDF]
Qiankun Fu, Yue Zhang, Jiangming Liu, and Meishan Zhang 13:00–14:00

Discourse representation tree structure (DRTS) parsing is a novel semantic parsing task which has been concerned most recently. State-of-the-art performance can be achieved by a neural sequence-to-sequence model, treating the tree construction as an incremental sequence generation problem. Structural information such as input syntax and the intermediate skeleton of the partial output has been ignored in the model, which could be potentially useful for the DRTS parsing. In this work, we propose a structural-aware model at both the encoder and decoder phase to integrate the structural information, where graph attention network (GAT) is exploited for effectively modeling. Experimental results on a benchmark dataset show that our proposed model is effective and can obtain the best performance in the literature.

Session 13B: Information Extraction-8

Bipartite Flat-Graph Network for Nested Named Entity Recognition [Website][PDF]
Ying Luo and Hai Zhao 13:00–14:00

In this paper, we propose a novel bipartite flat-graph network (BiFlaG) for nested named entity recognition (NER), which contains two subgraph modules: a flat NER module for outermost entities and a graph module for all the entities located in inner layers. Bidirectional LSTM (BiLSTM) and graph convolutional network (GCN) are adopted to jointly learn flat entities and their inner dependencies. Different from previous models, which only consider the unidirectional delivery of information from innermost layers to outer ones (or outside-to-inside), our model effectively captures the bidirectional interaction between them. We first use the entities recognized by the flat NER module to construct an entity graph, which is fed to the next graph module. The richer representation learned from graph module carries the dependencies of inner entities and can be exploited to improve outermost entity predictions. Experimental results on three standard nested NER datasets demonstrate that our BiFlaG outperforms previous state-of-the-art models.

FLAT: Chinese NER Using Flat-Lattice Transformer [Website][PDF]
Xiaonan Li, Hang Yan, Xipeng Qiu, and Xuanjing Huang 13:00–14:00

Recently, the character-word lattice structure has been proved to be effective for Chinese named entity recognition (NER) by incorporating the word information. However, since the lattice structure is complex and dynamic, the lattice-based models are hard to fully utilize the parallel computation of GPUs and usually have a low inference speed. In this paper, we propose FLAT: Flat-LAttice Transformer for Chinese NER, which converts the lattice structure into a flat structure consisting of spans. Each span corresponds to a character or latent word and its position in the original lattice. With the power of Transformer and well-designed position encoding, FLAT can fully leverage the lattice information and has an excellent parallel ability. Experiments on four datasets show FLAT outperforms other lexicon-based models in performance and efficiency.

Temporally-Informed Analysis of Named Entity Recognition [Website][PDF]
Shruti Rijhwani and Daniel Preotiuc-Pietro 13:00–14:00

Natural language processing models often have to make predictions on text data that evolves over time as a result of changes in language use or the information described in the text. However, evaluation results on existing data sets are seldom reported by taking the timestamp of the document into account. We analyze and propose methods that make better use of temporally-diverse training data, with a focus on the task of named entity recognition. To support these experiments, we introduce a novel data set of English tweets annotated with named entities. We empirically demonstrate the effect of temporal drift on performance, and how the temporal information of documents can be used to obtain better models compared to those that disregard temporal information. Our analysis gives insights into why this information is useful, in the hope of informing potential avenues of improvement for named entity recognition as well as other NLP tasks under similar experimental setups.

Towards Open Domain Event Trigger Identification using Adversarial Domain Adaptation [Website][PDF]
Aakanksha Naik and Carolyn Rose 13:00–14:00

We tackle the task of building supervised event trigger identification models which can generalize better across domains. Our work leverages the adversarial domain adaptation (ADA) framework to introduce domain-invariance. ADA uses adversarial training to construct representations that are predictive for trigger identification, but not predictive of the example's domain. It requires no labeled data from the target domain, making it completely unsupervised. Experiments with two domains (English literature and news) show that ADA leads to an average F1 score improvement of 3.9 on out-of-domain data. Our best performing model (BERT-A) reaches 44-49 F1 across both domains, using no labeled target data. Preliminary experiments reveal that finetuning on 1% labeled data, followed by self-training leads to substantial improvement, reaching 51.5 and 67.2 F1 on literature and news respectively.

Session 13B: Language Grounding to Vision, Robotics and Beyond-7

Aligned Dual Channel Graph Convolutional Network for Visual Question Answering [Website][PDF]
Qingbao Huang, Jielong Wei, Yi Cai, Changmeng Zheng, Junying Chen, Ho-fung Leung, and Qing Li
13:00–14:00

Visual question answering aims to answer the natural language question about a given image. Existing graph-based methods only focus on the relations between objects in an image and neglect the importance of the syntactic dependency relations between words in a question. To simultaneously capture the relations between objects in an image and the syntactic dependency relations between words in a question, we propose a novel dual channel graph convolutional network (DC-GCN) for better combining visual and textual advantages. The DC-GCN model consists of three parts: an I-GCN module to capture the relations between objects in an image, a Q-GCN module to capture the syntactic dependency relations between words in a question, and an attention alignment module to align image representations and question representations. Experimental results show that our model achieves comparable performance with the state-of-the-art approaches.

CompGuessWhat?!: A Multi-task Evaluation Framework for Grounded Language Learning [Website][PDF]
Alessandro Suglia, Ioannis Konstas, Andrea Vanzo, Emanuele Bastianelli, Desmond Elliott, Stella Frank, and Oliver Lemon 13:00–14:00

Approaches to Grounded Language Learning are commonly focused on a single task-based final performance measure which may not depend on desirable properties of the learned hidden representations, such as their ability to predict object attributes or generalize to unseen situations. To remedy this, we present GroLLA, an evaluation framework for Grounded Language Learning with Attributes based on three sub-tasks: 1) Goal-oriented evaluation; 2) Object attribute prediction evaluation; and 3) Zero-shot evaluation. We also propose a new dataset CompGuessWhat?! as an instance of this framework for evaluating the quality of learned neural representations, in particular with respect to attribute grounding. To this end, we extend the original GuessWhat?! dataset by including a semantic layer on top of the perceptual one. Specifically, we enrich the VisualGenome scene graphs associated with the GuessWhat?! images with several attributes from resources such as VISA and ImSitu. We then compare several hidden state representations from current state-of-the-art approaches to Grounded Language Learning. By using diagnostic classifiers, we show that current models' learned representations are not expressive enough to encode object attributes (average F1 of 44.27). In addition, they do not learn strategies nor representations that are robust enough to perform well when novel scenes or objects are involved in gameplay (zero-shot best accuracy 50.06%).

Cross-Modality Relevance for Reasoning on Language and Vision [Website][PDF]
Chen Zheng, Quan Guo, and Parisa Kordjamshidi 13:00–14:00

This work deals with the challenge of learning and reasoning over language and vision data for the related downstream tasks such as visual question answering (VQA) and natural language for visual reasoning (NLVR). We design a novel cross-modality relevance module that is used in an end-to-end framework to learn the relevance representation between components of various input modalities under the supervision of a target task, which is more generalizable to unobserved data compared to merely reshaping the original representation space. In addition to modeling the relevance between the textual entities and visual entities, we model the higher-order relevance between entity relations in the text and object relations in the image. Our proposed approach shows competitive performance on two different language and vision tasks using public benchmarks and improves the state-of-the-art published results. The learned alignments of input spaces and their relevance representations by NLVR task boost the training efficiency of VQA task.

Learning Web-based Procedures by Reasoning over Explanations and Demonstrations in Context [Website][PDF]
Shashank Srivastava, Oleksandr Polozov, Nebojsa Jojic, and Christopher Meek 13:00–14:00

We explore learning web-based tasks from a human teacher through natural language explanations and a single demonstration. Our approach investigates a new direction for semantic parsing that models explaining a demonstration in a context, rather than mapping explanations to demonstrations. By leveraging the idea of inverse semantics from program synthesis to reason backwards from observed demonstrations, we ensure that all considered interpretations are consistent with executable actions in any context, thus simplifying the problem of search over logical forms. We present a dataset of explanations paired with demonstrations for web-based tasks. Our methods show better task completion rates than a supervised semantic parsing baseline (40% relative improvement on average), and are competitive with simple exploration-and-demonstration based methods, while requiring no exploration of the environment. In learning to align explanations with demonstrations, basic properties of natural language syntax emerge as learned behavior. This is an interesting example of pragmatic language acquisition without any linguistic annotation.

Multi-agent Communication meets Natural Language: Synergies between Functional and Structural Language Learning [Website][PDF]
Angeliki Lazaridou, Anna Potapenko, and Olivier Tieleman 13:00–14:00

We present a method for combining multi-agent communication and traditional data-driven approaches to natural language learning, with an end goal of teaching agents to communicate with humans in natural language. Our starting point is a language model that has been trained on generic, not task-specific language data. We then place this model in a multi-agent self-play environment that generates task-specific rewards used to adapt or modulate the model, turning it into a task-conditional language model. We introduce a new way for combining the two types of learning based on the idea of reranking language model samples, and show that this method outperforms others in

communicating with humans in a visual referential communication task. Finally, we present a taxonomy of different types of language drift that can occur alongside a set of measures to detect them.

Multimodal Neural Graph Memory Networks for Visual Question Answering [Website] [PDF]
Mahmoud Khademi 13:00–14:00

We introduce a new neural network architecture, Multimodal Neural Graph Memory Networks (MN-GMN), for visual question answering. The MN-GMN uses graph structure with different region features as node attributes and applies a recently proposed powerful graph neural network model, Graph Network (GN), to reason about objects and their interactions in an image. The input module of the MN-GMN generates a set of visual features plus a set of encoded region-grounded captions (RGCs) for the image. The RGCs capture object attributes and their relationships. Two GNs are constructed from the input module using the visual features and encoded RGCs. Each node of the GNs iteratively computes a question-guided contextualized representation of the visual/textual information assigned to it. Then, to combine the information from both GNs, the nodes write the updated representations to an external spatial memory. The final states of the memory cells are fed into an answer module to predict an answer. Experiments show MN-GMN rivals the state-of-the-art models on Visual7W, VQA-v2.0, and CLEVR datasets.

Session 13B: Machine Translation-15

HAT: Hardware-Aware Transformers for Efficient Natural Language Processing [Website][PDF]
Hanrui Wang, Zhanghao Wu, Zhijian Liu, Han Cai, Ligeng Zhu, Chuang Gan, and Song Han 13:00–14:00

Transformers are ubiquitous in Natural Language Processing (NLP) tasks, but they are difficult to be deployed on hardware due to the intensive computation. To enable low-latency inference on resource-constrained hardware platforms, we propose to design Hardware-Aware Transformers (HAT) with neural architecture search. We first construct a large design space with arbitrary encoder-decoder attention and heterogeneous layers. Then we train a SuperTransformer that covers all candidates in the design space, and efficiently produces many SubTransformers with weight sharing. Finally, we perform an evolutionary search with a hardware latency constraint to find a specialized SubTransformer dedicated to run fast on the target hardware. Extensive experiments on four machine translation tasks demonstrate that HAT can discover efficient models for different hardware (CPU, GPU, IoT device). When running WMT'14 translation task on Raspberry Pi-4, HAT can achieve 3× speedup, 3.7× smaller size over baseline Transformer; 2.7× speedup, 3.6× smaller size over Evolved Transformer with 12,041× less search cost and no performance loss. HAT is open-sourced at https://github.com/mit-han-lab/hardware-aware-transformers.

Hard-Coded Gaussian Attention for Neural Machine Translation [Website][PDF]
Weiqiu You, Simeng Sun, and Mohit Iyyer 13:00–14:00

Recent work has questioned the importance of the Transformer's multi-headed attention for achieving high translation quality. We push further in this direction by developing a "hard-coded" attention variant without any learned parameters. Surprisingly, replacing all learned self-attention heads in the encoder and decoder with fixed, input-agnostic Gaussian distributions minimally impacts BLEU scores across four different language pairs. However, additionally, hard-coding cross attention (which connects the decoder to the encoder) significantly lowers BLEU, suggesting that it is more important than self-attention. Much of this BLEU drop can be recovered by adding just a single learned cross attention head to an otherwise hard-coded Transformer. Taken as a whole, our results offer insight into which components of the Transformer are actually important, which we hope will guide future work into the development of simpler and more efficient attention-based models.

In Neural Machine Translation, What Does Transfer Learning Transfer? [Website][PDF]
Alham Fikri Aji, Nikolay Bogoychev, Kenneth Heafield, and Rico Sennrich 13:00–14:00

Transfer learning improves quality for low-resource machine translation, but it is unclear what exactly it transfers. We perform several ablation studies that limit information transfer, then measure the quality impact across three language pairs to gain a black-box understanding of transfer learning. Word embeddings play an important role in transfer learning, particularly if they are properly aligned. Although transfer learning can be performed without embeddings, results are sub-optimal. In contrast, transferring only the embeddings but nothing else yields catastrophic results. We then investigate diagonal alignments with auto-encoders over real languages and randomly generated sequences, finding even randomly generated sequences as parents yield noticeable but smaller gains. Finally, transfer learning can eliminate the need for a warm-up phase when training transformer models in high resource language pairs.

Learning a Multi-Domain Curriculum for Neural Machine Translation [Website][PDF]
Wei Wang, Ye Tian, Jiquan Ngiam, Yinfei Yang, Isaac Caswell, and Zarana Parekh 13:00–14:00

Most data selection research in machine translation focuses on improving a single domain. We perform data selection for multiple domains at once. This is achieved by carefully introducing instance-level domain-relevance features and automatically constructing a training curriculum to gradually concentrate on multi-domain relevant and noise-reduced data batches. Both the choice of features and the use of curriculum are crucial for balancing and improving all domains, including out-of-domain. In large-scale experiments, the multi-domain curriculum simultaneously reaches or outperforms the individual performance and brings solid gains over no-curriculum training.

Reducing Gender Bias in Neural Machine Translation as a Domain Adaptation Problem [Website][PDF]
Danielle Saunders and Bill Byrne 13:00–14:00

Training data for NLP tasks often exhibits gender bias in that fewer sentences refer to women than to men. In Neural Machine Translation (NMT) gender bias has been shown to reduce translation quality, particularly when the target language has grammatical gender. The recent WinoMT challenge set allows us to measure this effect directly (Stanovsky et al, 2019) Ideally we would reduce system bias by simply debiasing all data prior to training, but achieving this effectively is itself a challenge. Rather than attempt to create a 'balanced' dataset, we use transfer learning on a small set of trusted, gender-balanced examples. This approach gives strong and consistent improvements in gender debiasing with much less computational cost than training from scratch. A known pitfall of transfer learning on new domains is 'catastrophic forgetting', which we address at adaptation and inference time. During adaptation we show that Elastic Weight Consolidation allows a performance trade-off between general translation quality and bias reduction. At inference time we propose a lattice-rescoring scheme which outperforms all systems evaluated in Stanovsky et al, 2019 on WinoMT with no degradation of general test set BLEU. We demonstrate our approach translating from English into three languages with varied linguistic properties and data availability.

Translationese as a Language in "Multilingual" NMT [Website][PDF]
Parker Riley, Isaac Caswell, Markus Freitag, and David Grangier 13:00–14:00

Machine translation has an undesirable propensity to produce "translationese" artifacts, which can lead to higher BLEU scores while being liked less by human raters. Motivated by this, we model translationese and original (i.e.

natural) text as separate languages in a multilingual model, and pose the question: can we perform zero-shot translation between original source text and original target text? There is no data with original source and original target, so we train a sentence-level classifier to distinguish translationese from original target text, and use this classifier to tag the training data for an NMT model. Using this technique we bias the model to produce more natural outputs at test time, yielding gains in human evaluation scores on both accuracy and fluency. Additionally, we demonstrate that it is possible to bias the model to produce translationese and game the BLEU score, increasing it while decreasing human-rated quality. We analyze these outputs using metrics measuring the degree of translationese, and present an analysis of the volatility of heuristic-based train-data tagging.

Uncertainty-Aware Curriculum Learning for Neural Machine Translation [Website] [PDF]
Yikai Zhou, Baosong Yang, Derek F. Wong, Yu Wan, and Lidia S. Chao 13:00–14:00

Neural machine translation (NMT) has proven to be facilitated by curriculum learning which presents examples in an easy-to-hard order at different training stages. The keys lie in the assessment of data difficulty and model competence. We propose uncertainty-aware curriculum learning, which is motivated by the intuition that: 1) the higher the uncertainty in a translation pair, the more complex and rarer the information it contains; and 2) the end of the decline in model uncertainty indicates the completeness of current training stage. Specifically, we serve cross-entropy of an example as its data difficulty and exploit the variance of distributions over the weights of the network to present the model uncertainty. Extensive experiments on various translation tasks reveal that our approach outperforms the strong baseline and related methods on both translation quality and convergence speed. Quantitative analyses reveal that the proposed strategy offers NMT the ability to automatically govern its learning schedule.

Unsupervised Domain Clusters in Pretrained Language Models [Website] [PDF]
Roee Aharoni and Yoav Goldberg 13:00–14:00

The notion of "in-domain data" in NLP is often over-simplistic and vague, as textual data varies in many nuanced linguistic aspects such as topic, style or level of formality. In addition, domain labels are many times unavailable, making it challenging to build domain-specific systems. We show that massive pre-trained language models implicitly learn sentence representations that cluster by domains without supervision – suggesting a simple data-driven definition of domains in textual data. We harness this property and propose domain data selection methods based on such models, which require only a small set of in-domain monolingual data. We evaluate our data selection methods for neural machine translation across five diverse domains, where they outperform an established approach as measured by both BLEU and precision and recall with respect to an oracle selection.

Using Context in Neural Machine Translation Training Objectives [Website] [PDF]
Danielle Saunders, Felix Stahlberg, and Bill Byrne 13:00–14:00

We present Neural Machine Translation (NMT) training using document-level metrics with batch-level documents. Previous sequence-objective approaches to NMT training focus exclusively on sentence-level metrics like sentence BLEU which do not correspond to the desired evaluation metric, typically document BLEU. Meanwhile research into document-level NMT training focuses on data or model architecture rather than training procedure. We find that each of these lines of research has a clear space in it for the other, and propose merging them with a scheme that allows a document-level evaluation metric to be used in the NMT training objective. We first sample pseudo-documents from sentence samples. We then approximate the expected document BLEU gradient with Monte Carlo sampling for use as a cost function in Minimum Risk Training (MRT). This two-level sampling procedure gives NMT performance gains over sequence MRT and maximum-likelihood training. We demonstrate that training is more robust for document-level metrics than with sequence metrics. We further demonstrate improvements on NMT with TER and Grammatical Error Correction (GEC) using GLEU, both metrics used at the document level for evaluations.

Variational Neural Machine Translation with Normalizing Flows [Website] [PDF]
Hendra Setiawan, Matthias Sperber, Udhyakumar Nallasamy, and Matthias Paulik 13:00–14:00

Variational Neural Machine Translation (VNMT) is an attractive framework for modeling the generation of target translations, conditioned not only on the source sentence but also on some latent random variables. The latent variable modeling may introduce useful statistical dependencies that can improve translation accuracy. Unfortunately, learning informative latent variables is non-trivial, as the latent space can be prohibitively large, and the latent codes are prone to be ignored by many translation models at training time. Previous works impose strong assumptions on the distribution of the latent code and limit the choice of the NMT architecture. In this paper, we propose to apply the VNMT framework to the state-of-the-art Transformer and introduce a more flexible approximate posterior based on normalizing flows. We demonstrate the efficacy of our proposal under both in-domain and out-of-domain conditions, significantly outperforming strong baselines.

Session 13B: Phonology, Morphology and Word Segmentation-5

2kenize: Tying Subword Sequences for Chinese Script Conversion [Website][PDF]
Pranav A and Isabelle Augenstein 13:00–14:00

Simplified Chinese to Traditional Chinese character conversion is a common preprocessing step in Chinese NLP. Despite this, current approaches have insufficient performance because they do not take into account that a simplified Chinese character can correspond to multiple traditional characters. Here, we propose a model that can disambiguate between mappings and convert between the two scripts. The model is based on subword segmentation, two language models, as well as a method for mapping between subword sequences. We further construct benchmark datasets for topic classification and script conversion. Our proposed method outperforms previous Chinese Character conversion approaches by 6 points in accuracy. These results are further confirmed in a downstream application, where 2kenize is used to convert pretraining dataset for topic classification. An error analysis reveals that our method's particular strengths are in dealing with code mixing and named entities.

[TACL] Phonotactic Complexity and its Trade-offs [Website][PDF]
Tiago Pimentel, Brian Roark, and Ryan D. Cotterell 13:00–14:00

We present methods for calculating a measure of phonotactic complexity—bits per phoneme—that permits a straightforward cross-linguistic comparison. When given a word, represented as a sequence of phonemic segments such as symbols in the international phonetic alphabet, and a statistical model trained on a sample of word types from the language, we can approximately measure bits per phoneme using the negative log-probability of that word under the model. This simple measure allows us to compare the entropy across languages, giving insight into how complex a language's phonotactics are. Using a collection of 1016 basic concept words across 106 languages, we demonstrate a very strong negative correlation of -0.74 between bits per phoneme and the average length of words.

Predicting the Growth of Morphological Families from Social and Linguistic Factors [Website][PDF]
Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze 13:00–14:00

We present the first study that examines the evolution of morphological families, i.e., sets of morphologically related words such as "trump", "antitrumpism", and "detrumpify", in social media. We introduce the novel task of Morphological Family Expansion Prediction (MFEP) as predicting the increase in the size of a morphological family. We create a ten-year Reddit corpus as a benchmark for MFEP and evaluate a number of baselines on this benchmark. Our experiments demonstrate very good performance on MFEP.

Semi-supervised Contextual Historical Text Normalization [Website][PDF]
Peter Makarov and Simon Clematide 13:00–14:00

Historical text normalization, the task of mapping historical word forms to their modern counterparts, has recently attracted a lot of interest (Bollmann, 2019; Tang et al., 2018; Lusetti et al., 2018; Bollmann et al., 2018;Robertson and Goldwater, 2018; Bollmannet al., 2017; Korchagina, 2017). Yet, virtually all approaches suffer from the two limitations: 1) They consider a fully supervised setup, often with impractically large manually normalized datasets; 2) Normalization happens on words in isolation. By utilizing a simple generative normalization model and obtaining powerful contextualization from the target-side language model, we train accurate models with unlabeled historical data. In realistic training scenarios, our approach often leads to reduction in manually normalized data at the same accuracy levels.

The Paradigm Discovery Problem [Website][PDF]
Alexander Erdmann, Micha Elsner, Shijie Wu, Ryan Cotterell, and Nizar Habash 13:00–14:00

This work treats the paradigm discovery problem (PDP), the task of learning an inflectional morphological system from unannotated sentences. We formalize the PDP and develop evaluation metrics for judging systems. Using currently available resources, we construct datasets for the task. We also devise a heuristic benchmark for the PDP and report empirical results on five diverse languages. Our benchmark system first makes use of word embeddings and string similarity to cluster forms by cell and by paradigm. Then, we bootstrap a neural transducer on top of the clustered data to predict words to realize the empty paradigm slots. An error analysis of our system suggests clustering by cell across different inflection classes is the most pressing challenge for future work.

Supervised Grapheme-to-Phoneme Conversion of Orthographic Schwas in Hindi and Punjabi [Website][PDF]
Aryaman Arora, Luke Gessler, and Nathan Schneider 13:00–14:00

Hindi grapheme-to-phoneme (G2P) conversion is mostly trivial, with one exception: whether a schwa represented in the orthography is pronounced or unpronounced (deleted). Previous work has attempted to predict schwa deletion in a rule-based fashion using prosodic or phonetic analysis. We present the first statistical schwa deletion classifier for Hindi, which relies solely on the orthography as the input and outperforms previous approaches. We trained our model on a newly-compiled pronunciation lexicon extracted from various online dictionaries. Our best Hindi model achieves state of the art performance, and also achieves good performance on a closely related language, Punjabi, without modification.

Session 13B: Question Answering-11

ClarQ: A large-scale and diverse dataset for Clarification Question Generation [Website][PDF]
Vaibhav Kumar and Alan W Black 13:00–14:00

Question answering and conversational systems are often baffled and need help clarifying certain ambiguities. However, limitations of existing datasets hinder the development of large-scale models capable of generating and utilising clarification questions. In order to overcome these limitations, we devise a novel bootstrapping framework (based on self-supervision) that assists in the creation of a diverse, large-scale dataset of clarification questions based on post-comment tuples extracted from stackexchange. The framework utilises a neural network based architecture for classifying clarification questions. It is a two-step method where the first aims to increase the precision of the classifier and second aims to increase its recall. We quantitatively demonstrate the utility of the newly created dataset by applying it to the downstream task of question-answering. The final dataset, ClarQ, consists of ~2M examples distributed across 173 domains of stackexchange. We release this dataset in order to foster research into the field of clarification question generation with the larger goal of enhancing dialog and question answering systems.

DoQA - Accessing Domain-Specific FAQs via Conversational QA [Website][PDF]
Jon Ander Campos, Arantxa Otegi, Aitor Soroa, Jan Deriu, Mark Cieliebak, and Eneko Agirre 13:00–14:00

The goal of this work is to build conversational Question Answering (QA) interfaces for the large body of domain-specific information available in FAQ sites. We present DoQA, a dataset with 2,437 dialogues and 10,917 QA pairs. The dialogues are collected from three Stack Exchange sites using the Wizard of Oz method with crowdsourcing. Compared to previous work, DoQA comprises well-defined information needs, leading to more coherent and natural conversations with less factoid questions and is multi-domain. In addition, we introduce a more realistic information retrieval (IR) scenario where the system needs to find the answer in any of the FAQ documents. The results of an existing, strong, system show that, thanks to transfer learning from a Wikipedia QA dataset and fine tuning on a single FAQ domain, it is possible to build high quality conversational QA systems for FAQs without in-domain training data. The good results carry over into the more challenging IR scenario. In both cases, there is still ample room for improvement, as indicated by the higher human upperbound.

Document Modeling with Graph Attention Networks for Multi-grained Machine Reading Comprehension [Website][PDF]
Bo Zheng, Haoyang Wen, Yaobo Liang, Nan Duan, Wanxiang Che, Daxin Jiang, Ming Zhou, and Ting Liu 13:00–14:00

Natural Questions is a new challenging machine reading comprehension benchmark with two-grained answers, which are a long answer (typically a paragraph) and a short answer (one or more entities inside the long answer). Despite the effectiveness of existing methods on this benchmark, they treat these two sub-tasks individually during training while ignoring their dependencies. To address this issue, we present a novel multi-grained machine reading comprehension framework that focuses on modeling documents at their hierarchical nature, which are different levels of granularity: documents, paragraphs, sentences, and tokens. We utilize graph attention networks to obtain different levels of representations so that they can be learned simultaneously. The long and short answers can be extracted from paragraph-level representation and token-level representation, respectively. In this way, we can model the dependencies between the two-grained answers to provide evidence for each other. We jointly train the two sub-tasks, and our experiments show that our approach significantly outperforms previous systems at both long and short answer criteria.

Low-Resource Generation of Multi-hop Reasoning Questions [Website][PDF]
Jianxing Yu, Wei Liu, Shuang Qiu, Qinliang Su, Kai Wang, Xiaojun Quan, and Jian Yin 13:00–14:00

This paper focuses on generating multi-hop reasoning questions from the raw text in a low resource circumstance. Such questions have to be syntactically valid and need to logically correlate with the answers by deducing over multiple relations on several sentences in the text. Specifically, we first build a multi-hop generation model and guide it to satisfy the logical rationality by the reasoning chain extracted from a given text. Since the labeled data is limited and insufficient for training, we propose to learn the model with the help of a large scale of unlabeled data that is much easier to obtain. Such data contains rich expressive forms of the questions with structural patterns on syntax and semantics. These patterns can be estimated by the neural hidden semi-Markov model using latent variables. With latent patterns as a prior, we can regularize the generation model and produce the optimal results. Experimental results on the HotpotQA data set demonstrate the effectiveness of our model. Moreover, we apply the generated results to the task of machine reading comprehension and achieve significant performance improvements.

MLQA: Evaluating Cross-lingual Extractive Question Answering [Website][PDF]
Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk 13:00–14:00

Question answering (QA) models have shown rapid progress enabled by the availability of large, high-quality benchmark datasets. Such annotated datasets are difficult and costly to collect, and rarely exist in languages other than English, making building QA systems that work well in other languages challenging. In order to develop such systems, it is crucial to invest in high quality multilingual evaluation benchmarks to measure progress. We present MLQA, a multi-way aligned extractive QA evaluation benchmark intended to spur research in this area. MLQA contains QA instances in 7 languages, English, Arabic, German, Spanish, Hindi, Vietnamese and Simplified Chinese. MLQA has over 12K instances in English and 5K in each other language, with each instance parallel between 4 languages on average. We evaluate state-of-the-art cross-lingual models and machine-translation-based baselines on MLQA. In all cases, transfer results are shown to be significantly behind training-language performance.

Multi-source Meta Transfer for Low Resource Multiple-Choice Question Answering [Website][PDF]
Ming Yan, Hao Zhang, Di Jin, and Joey Tianyi Zhou 13:00–14:00

Multiple-choice question answering (MCQA) is one of the most challenging tasks in machine reading comprehension since it requires more advanced reading comprehension skills such as logical reasoning, summarization, and arithmetic operations. Unfortunately, most existing MCQA datasets are small in size, which increases the difficulty of model learning and generalization. To address this challenge, we propose a multi-source meta transfer (MMT) for low-resource MCQA. In this framework, we first extend meta learning by incorporating multiple training sources to learn a generalized feature representation across domains. To bridge the distribution gap between training sources and the target, we further introduce the meta transfer that can be integrated into the multi-source meta training. More importantly, the proposed MMT is independent of backbone language models. Extensive experiments demonstrate the superiority of MMT over state-of-the-arts, and continuous improvements can be achieved on different backbone networks on both supervised and unsupervised domain adaptation settings.

R4C: A Benchmark for Evaluating RC Systems to Get the Right Answer for the Right Reason [Website][PDF]
Naoya Inoue, Pontus Stenetorp, and Kentaro Inui 13:00–14:00

Recent studies have revealed that reading comprehension (RC) systems learn to exploit annotation artifacts and other biases in current datasets. This prevents the community from reliably measuring the progress of RC systems. To address this issue, we introduce R4C, a new task for evaluating RC systems' internal reasoning. R4C requires giving not only answers but also derivations: explanations that justify predicted answers. We present a reliable, crowdsourced framework for scalably annotating RC datasets with derivations. We create and publicly release the R4C dataset, the first, quality-assured dataset consisting of 4.6k questions, each of which is annotated with 3 reference derivations (i.e. 13.8k derivations). Experiments show that our automatic evaluation metrics using multiple reference derivations are reliable, and that R4C assesses different skills from an existing benchmark.

Recurrent Chunking Mechanisms for Long-Text Machine Reading Comprehension [Website][PDF]
Hongyu Gong, Yelong Shen, Dian Yu, Jianshu Chen, and Dong Yu 13:00–14:00

In this paper, we study machine reading comprehension (MRC) on long texts: where a model takes as inputs a lengthy document and a query, extracts a text span from the document as an answer. State-of-the-art models (e.g., BERT) tend to use a stack of transformer layers that are pre-trained from a large number of unlabeled language corpora to encode the joint contextual information of query and document. However, these transformer models can only take as input a fixed-length (e.g., 512) text. To deal with even longer text inputs, previous approaches usually chunk them into *equally-spaced* segments and predict answers based on each segment independently without considering the information from other segments. As a result, they may form segments that fail to cover complete answers or retain insufficient contexts around the correct answer required for question answering. Moreover, they are less capable of answering questions that need cross-segment information. We propose to let a model learn to chunk in a more flexible way via reinforcement learning: a model can decide the next segment that it wants to process in either direction. We also apply recurrent mechanisms to enable information to flow across segments. Experiments on three MRC tasks – CoQA, QuAC, and TriviaQA – demonstrate the effectiveness of our proposed recurrent chunking mechanisms: we can obtain segments that are more likely to contain complete answers and at the same time provide sufficient contexts around the ground truth answers for better predictions.

Session 13B: Student Research Workshop

AraDIC: Arabic Document Classification Using Image-Based Character Embeddings and Class-Balanced Loss [Website][PDF]
Mahmoud Daif, Shunsuke Kitada, and Hitoshi Iyatomi 13:00–14:00

Classical and some deep learning techniques for Arabic text classification often depend on complex morphological analysis, word segmentation, and hand-crafted feature engineering. These could be eliminated by using character-level features. We propose a novel end-to-end Arabic document classification framework, Arabic document image-based classifier (AraDIC), inspired by the work on image-based character embeddings. AraDIC consists of an image-based character encoder and a classifier. They are trained in an end-to-end fashion using the class balanced loss to deal with the long-tailed data distribution problem. To evaluate the effectiveness of AraDIC, we created and published two datasets, the Arabic Wikipedia title (AWT) dataset and the Arabic poetry (AraP) dataset. To the best of our knowledge, this is the first image-based character embedding framework addressing the problem of Arabic text classification. We also present the first deep learning-based text classifier widely evaluated on modern standard Arabic, colloquial Arabic, and Classical Arabic. AraDIC shows performance improvement over classical and deep learning baselines by 12.29% and 23.05% for the micro and macro F-score, respectively.

Understanding Points of Correspondence between Sentences for Abstractive Summarization [Website][PDF]
Logan Lebanoff, John Muchovej, Franck Dernoncourt, Doo Soon Kim, Lidan Wang, Walter Chang, and Fei Liu 13:00–14:00

Fusing sentences containing disparate content is a remarkable human ability that helps create informative and succinct summaries. Such a simple task for humans has remained challenging for modern abstractive summarizers, substantially restricting their applicability in real-world scenarios. In this paper, we present an investigation into fusing sentences drawn from a document by introducing the notion of points of correspondence, which are cohesive devices that tie any two sentences together into a coherent text. The types of points of correspondence are delineated by text cohesion theory, covering pronominal and nominal referencing, repetition and beyond. We create a dataset containing the documents, source and fusion sentences, and human annotations of points of correspondence between sentences. Our dataset bridges the gap between coreference resolution and summarization. It is publicly shared to serve as a basis for future work to measure the success of sentence fusion systems.

Noise-Based Augmentation Techniques for Emotion Datasets: What do we Recommend? [Website]
Mimansa Jaiswal and Emily Mower Provost 13:00–14:00

Emotion recognition systems are widely used for many downstream applications such as mental health monitoring, educational problems diagnosis, hate speech classification and targeted advertising. Yet, these systems are generally trained on audio or multimodal datasets collected in a laboratory environment.While acoustically different, they are generally free of major environmental noises. The result is that systems trained on these datasets falter when presented with noisy data, even when that noise doesn't affect the human perception of emotions. In this work, we use multiple categories of environmental and synthetic noises to generate black box adversarial examples and use these noises to modify the samples in the IEMOCAP dataset. We evaluate how both human and machine emotion perception changes when noise is introduced. We find that the trained state-of-the-art models fail to classify even moderately noisy samples that humans usually have no trouble comprehend-ing, demonstrating the brittleness of these systems in real world conditions.

Logical Inferences with Comparatives and Generalized Quantifiers [Website][PDF]
Izumi Haruta, Koji Mineshima, and Daisuke Bekki 13:00–14:00

Comparative constructions pose a challenge in Natural Language Inference (NLI), which is the task of determining whether a text entails a hypothesis. Comparatives are structurally complex in that they interact with other linguistic phenomena such as quantifiers, numerals, and lexical antonyms. In formal semantics, there is a rich body of work on comparatives and gradable expressions using the notion of degree. However, a logical inference system for comparatives has not been sufficiently developed for use in the NLI task. In this paper, we present a compositional semantics that maps various comparative constructions in English to semantic representations via Combinatory Categorial Grammar (CCG) parsers and combine it with an inference system based on automated theorem proving. We evaluate our system on three NLI datasets that contain complex logical inferences with comparatives, generalized quantifiers, and numerals. We show that the system outperforms previous logic-based systems as well as recent deep learning-based models.

Session 13B: Theme-5

A Call for More Rigor in Unsupervised Cross-lingual Learning [Website][PDF]
Mikel Artetxe, Sebastian Ruder, Dani Yogatama, Gorka Labaka, and Eneko Agirre 13:00–14:00

We review motivations, definition, approaches, and methodology for unsupervised cross-lingual learning and call for a more rigorous position in each of them. An existing rationale for such research is based on the lack of parallel data for many of the world's languages. However, we argue that a scenario without any parallel data and abundant monolingual data is unrealistic in practice. We also discuss different training signals that have been used in previous work, which depart from the pure unsupervised setting. We then describe common methodological issues in tuning and evaluation of unsupervised cross-lingual models and present best practices. Finally, we provide a unified outlook for different types of research in this area (i.e., cross-lingual word embeddings, deep multilingual pretraining, and unsupervised machine translation) and argue for comparable evaluation of these models.

A Tale of a Probe and a Parser [Website][PDF]
Rowan Hall Maudslay, Josef Valvoda, Tiago Pimentel, Adina Williams, and Ryan Cotterell 13:00–14:00

Measuring what linguistic information is encoded in neural models of language has become popular in NLP. Researchers approach this enterprise by training "probes"—supervised models designed to extract linguistic structure from another model's output. One such probe is the structural probe (Hewitt and Manning, 2019), designed to quantify the extent to which syntactic information is encoded in contextualised word representations. The structural probe has a novel design, unattested in the parsing literature, the precise benefit of which is not immediately obvious. To explore whether syntactic probes would do better to make use of existing techniques, we compare the structural probe to a more traditional parser with an identical lightweight parameterisation. The parser outperforms structural probe on UUAS in seven of nine analysed languages, often by a substantial amount (e.g. by 11.1 points in English). Under a second less common metric, however, there is the opposite trend—the structural probe outperforms the parser. This begs the question: which metric should we prefer?

Are we Estimating or Guesstimating Translation Quality? [Website][PDF]
Shuo Sun, Francisco Guzmán, and Lucia Specia 13:00–14:00

Recent advances in pre-trained multilingual language models lead to state-of-the-art results on the task of quality estimation (QE) for machine translation. A carefully engineered ensemble of such models won the QE shared task at WMT19. Our in-depth analysis, however, shows that the success of using pre-trained language models for QE is over-estimated due to three issues we observed in current QE datasets: (i) The distributions of quality scores are imbalanced and skewed towards good quality scores; (iii) QE models can perform well on these datasets while looking at only source or translated sentences; (iii) They contain statistical artifacts that correlate well with human-annotated QE labels. Our findings suggest that although QE models might capture fluency of translated sentences and complexity of source sentences, they cannot model adequacy of translations effectively.

Automated Evaluation of Writing — 50 Years and Counting [Website][PDF]
Beata Beigman Klebanov and Nitin Madnani 13:00–14:00

In this theme paper, we focus on Automated Writing Evaluation (AWE), using Ellis Page's seminal 1966 paper to frame the presentation. We discuss some of the current frontiers in the field and offer some thoughts on the emergent uses of this technology.

From SPMRL to NMRL: What Did We Learn (and Unlearn) in a Decade of Parsing Morphologically-Rich Languages (MRLs)? [Website][PDF]
Reut Tsarfaty, Dan Bareket, Stav Klein, and Amit Seker 13:00–14:00

It has been exactly a decade since the first establishment of SPMRL, a research initiative unifying multiple research efforts to address the peculiar challenges of Statistical Parsing for Morphologically-Rich Languages (MRLs). Here we reflect on parsing MRLs in that decade, highlight the solutions and lessons learned for the architectural, modeling and lexical challenges in the pre-neural era, and argue that similar challenges re-emerge in neural architectures for MRLs. We then aim to offer a climax, suggesting that incorporating symbolic ideas proposed in SPMRL terms into nowadays neural architectures has the potential to push NLP for MRLs to a new level. We sketch a strategies for designing Neural Models for MRLs (NMRL), and showcase preliminary support for these strategies via investigating the task of multi-tagging in Hebrew, a morphologically-rich, high-fusion, language.

Language (Re)modelling: Towards Embodied Language Understanding [Website][PDF]
Ronen Tamari, Chen Shani, Tom Hope, Miriam R L Petruck, Omri Abend, and Dafna Shahaf 13:00–14:00

While natural language understanding (NLU) is advancing rapidly, today's technology differs from human-like language understanding in fundamental ways, notably in its inferior efficiency, interpretability, and generalization. This work proposes an approach to representation and learning based on the tenets of embodied cognitive linguistics (ECL). According to ECL, natural language is inherently executable (like programming languages), driven by mental simulation and metaphoric mappings over hierarchical compositions of structures and schemata learned through embodied interaction. This position paper argues that the use of grounding by metaphoric reasoning and simulation will greatly benefit NLU systems, and proposes a system architecture along with a roadmap towards realizing this vision.

Negated and Misprimed Probes for Pretrained Language Models: Birds Can Talk, But Cannot Fly [Website][PDF]
Nora Kassner and Hinrich Schütze 13:00–14:00

Building on Petroni et al. 2019, we propose two new probing tasks analyzing factual knowledge stored in Pretrained Language Models (PLMs). (1) Negation. We find that PLMs do not distinguish between negated ("Birds cannot [MASK]") and non-negated ("Birds can [MASK]") cloze questions. (2) Mispriming. Inspired by priming methods in human psychology, we add "misprimes" to cloze questions ("Talk? Birds can [MASK]"). We find that PLMs are easily distracted by misprimes. These results suggest that PLMs still have a long way to go to adequately learn human-like factual knowledge.

On Forgetting to Cite Older Papers: An Analysis of the ACL Anthology [Website][PDF]
Marcel Bollmann and Desmond Elliott 13:00–14:00

The field of natural language processing is experiencing a period of unprecedented growth, and with it a surge of published papers. This represents an opportunity for us to take stock of how we cite the work of other researchers, and whether this growth comes at the expense of "forgetting" about older literature. In this paper, we address this question through bibliographic analysis. By looking at the age of outgoing citations in papers published at selected ACL venues between 2010 and 2019, we find that there is indeed a tendency for recent papers to cite more recent work, but the rate at which papers older than 15 years are cited has remained relatively stable.

Returning the N to NLP: Towards Contextually Personalized Classification Models [Website][PDF]
Lucie Flek 13:00–14:00

Most NLP models today treat language as universal, even though socio- and psycholingustic research shows that the communicated message is influenced by the characteristics of the speaker as well as the target audience. This paper surveys the landscape of personalization in natural language processing and related fields, and offers a path forward to mitigate the decades of deviation of the NLP tools from sociolingustic findings, allowing to flexibly process the "natural" language of each user rather than enforcing a uniform NLP treatment. It outlines a possible direction to incorporate these aspects into neural NLP models by means of socially contextual personalization, and proposes to shift the focus of our evaluation strategies accordingly.

Speech Translation and the End-to-End Promise: Taking Stock of Where We Are [Website][PDF]
Matthias Sperber and Matthias Paulik 13:00–14:00

Over its three decade history, speech translation has experienced several shifts in its primary research themes; moving from loosely coupled cascades of speech recognition and machine translation, to exploring questions of tight coupling, and finally to end-to-end models that have recently attracted much attention. This paper provides a brief survey of these developments, along with a discussion of the main challenges of traditional approaches which stem from committing to intermediate representations from the speech recognizer, and from training cascaded models separately towards different objectives. Recent end-to-end modeling techniques promise a principled way of overcoming these issues by allowing joint training of all model components and removing the need for explicit intermediate representations. However, a closer look reveals that many end-to-end models fall short of solving these issues, due to compromises made to address data scarcity. This paper provides a unifying categorization and nomenclature that covers both traditional and recent approaches and that may help researchers by highlighting both trade-offs and open research questions.

To Test Machine Comprehension, Start by Defining Comprehension [Website][PDF]
Jesse Dunietz, Greg Burnham, Akash Bharadwaj, Owen Rambow, Jennifer Chu-Carroll, and Dave Ferrucci 13:00–14:00

Many tasks aim to measure machine reading comprehension (MRC), often focusing on question types presumed to be difficult. Rarely, however, do task designers start by considering what systems should in fact comprehend. In this paper we make two key contributions. First, we argue that existing approaches do not adequately define comprehension; they are too unsystematic about what content is tested. Second, we present a detailed definition of comprehension—a "Template of Understanding"—for a widely useful class of texts, namely short narratives. We then conduct an experiment that strongly suggests existing systems are not up to the task of narrative understanding as we define it.

What Question Answering can Learn from Trivia Nerds [Website][PDF]
Jordan Boyd-Graber and Benjamin Börschinger 13:00–14:00

In addition to the traditional task of machines answering questions, question answering (QA) research creates interesting, challenging questions that help systems how to answer questions and reveal the best systems. We argue that creating a QA dataset—and the ubiquitous leaderboard that goes with it—closely resembles running a trivia tournament: you write questions, have agents (either humans or machines) answer the questions, and declare a winner. However, the research community has ignored the hard-learned lessons from decades of the trivia community creating vibrant, fair, and effective question answering competitions. After detailing problems with existing QA datasets, we outline the key lessons—removing ambiguity, discriminating skill, and adjudicating disputes—that can transfer to QA research and how they might be implemented.

What are the Goals of Distributional Semantics? [Website][PDF]
Guy Emerson 13:00–14:00

Distributional semantic models have become a mainstay in NLP, providing useful features for downstream tasks. However, assessing long-term progress requires explicit long-term goals. In this paper, I take a broad linguistic perspective, looking at how well current models can deal with various semantic challenges. Given stark differences between models proposed in different subfields, a broad perspective is needed to see how we could integrate them. I conclude that, while linguistic insights can guide the design of model architectures, future progress will require balancing the often conflicting demands of linguistic expressiveness and computational tractability.

Gender Gap in Natural Language Processing Research: Disparities in Authorship and Citations [Website][PDF]

Saif M. Mohammad 13:00–14:00

Disparities in authorship and citations across genders can have substantial adverse consequences not just on the disadvantaged gender, but also on the field of study as a whole. In this work, we examine female first author percentages and the citations to their papers in Natural Language Processing. We find that only about 29% of first authors are female and only about 25% of last authors are female. Notably, this percentage has not improved since the mid 2000s. We also show that, on average, female first authors are cited less than male first authors, even when controlling for experience and area of research. We hope that recording citation and participation gaps across demographic groups will improve awareness of gender gaps and encourage more inclusiveness and fairness in research.

Demo Session 3C

Time: *13:30–14:15*

Embedding-based Scientific Literature Discovery in a Text Editor Application [Website][PDF]
Onur Gökçe, Jonathan Prada, Nikola I. Nikolov, Nianlong Gu, and Richard H.R. Hahnloser

Each claim in a research paper requires all relevant prior knowledge to be discovered, assimilated, and appropriately cited. However, despite the availability of powerful search engines and sophisticated text editing software, discovering relevant papers and integrating the knowledge into a manuscript remain complex tasks associated with high cognitive load. To define comprehensive search queries requires strong motivation from authors, irrespective of their familiarity with the research field. Moreover, switching between independent applications for literature discovery, bibliography management, reading papers, and writing text burdens authors further and interrupts their creative process. Here, we present a web application that combines text editing and literature discovery in an interactive user interface. The application is equipped with a search engine that couples Boolean keyword filtering with nearest neighbor search over text embeddings, providing a discovery experience tuned to an author's manuscript and his interests. Our application aims to take a step towards more enjoyable and effortless academic writing. The demo of the application (https://SciEditorDemo2020.herokuapp.com) and a short video tutorial (https://youtu.be/pkdVU60IcRc) are available online.

Penman: An Open-Source Library and Tool for AMR Graphs [Website][PDF]
Michael Wayne Goodman

Abstract Meaning Representation (AMR) (Banarescu et al., 2013) is a framework for semantic dependencies that encodes its rooted and directed acyclic graphs in a format called PENMAN notation. The format is simple enough that users of AMR data often write small scripts or libraries for parsing it into an internal graph representation, but there is enough complexity that these users could benefit from a more sophisticated and well-tested solution. The open-source Python library Penman provides a robust parser, functions for graph inspection and manipulation, and functions for formatting graphs into PENMAN notation. Many functions are also available in a command-line tool, thus extending its utility to non-Python setups.

Demo Session 4A

Time: *17:00–17:45*

Nakdan: Professional Hebrew Diacritizer [Website][PDF]
Avi Shmidman, Shaltiel Shmidman, Moshe Koppel, and Yoav Goldberg

We present a system for automatic diacritization of Hebrew Text. The system combines modern neural models with carefully curated declarative linguistic knowledge and comprehensive manually constructed tables and dictionaries. Besides providing state of the art diacritization accuracy, the system also supports an interface for manual editing and correction of the automatic output, and has several features which make it particularly useful for preparation of scientific editions of historical Hebrew texts. The system supports Modern Hebrew, Rabbinic Hebrew and Poetic Hebrew. The system is freely accessible for all use at http://nakdanpro.dicta.org.il

SUPP.AI: finding evidence for supplement-drug interactions [Website][PDF]
Lucy Wang, Oyvind Tafjord, Arman Cohan, Sarthak Jain, Sam Skjonsberg, Carissa Schoenick, Nick Botner, and Waleed Ammar

Dietary supplements are used by a large portion of the population, but information on their pharmacologic interactions is incomplete. To address this challenge, we present SUPP.AI, an application for browsing evidence of supplement-drug interactions (SDIs) extracted from the biomedical literature. We train a model to automatically extract supplement information and identify such interactions from the scientific literature. To address the lack of labeled data for SDI identification, we use labels of the closely related task of identifying drug-drug interactions (DDIs) for supervision. We fine-tune the contextualized word representations of the RoBERTa language model using labeled DDI data, and apply the fine-tuned model to identify supplement interactions. We extract 195k evidence sentences from 22M articles (P=0.82, R=0.58, F1=0.68) for 60k interactions. We create the SUPP.AI application for users to search evidence sentences extracted by our model. SUPP.AI is an attempt to close the information gap on dietary supplements by making up-to-date evidence on SDIs more discoverable for researchers, clinicians, and consumers. An informational video on how to use SUPP.AI is available at: https://youtu.be/dR0ucKdORwc

Keynote Address: Josh Tenenbaum

Cognitive and computational building blocks for more human-like language in machines

Abstract: Humans learn language building on more basic conceptual and computational resources that we can already see precursors of in infancy. These include capacities for causal reasoning, symbolic rule formation, rapid abstraction, and commonsense representations of events in terms of objects, agents and their interactions. I will talk about steps towards capturing these abilities in engineering terms, using tools from hierarchical Bayesian models, probabilistic programs, program induction, and neuro-symbolic architectures. I will show examples of how these tools have been applied in both cognitive science and AI contexts, and point to ways they might be useful in building more human-like language, learning and reasoning in machines.

Biography: Josh Tenenbaum is Professor of Computational Cognitive Science at MIT in the Department of Brain and Cognitive Sciences, the Computer Science and Artificial Intelligence Laboratory (CSAIL) and the Center for Brains, Minds and Machines (CBMM). He received his PhD from MIT in 1999, and taught at Stanford from 1999 to 2002. His long-term goal is to reverse-engineer intelligence in the human mind and brain, and use these insights to engineer more human-like machine intelligence. His current research focuses on the development of common sense in children and machines, the neural basis of common sense, and models of learning as Bayesian program synthesis. His work has been published in Science, Nature, PNAS, and many other leading journals, and recognized with awards at conferences in Cognitive Science, Computer Vision, Neural Information Processing Systems, Reinforcement Learning and Decision Making, and Robotics. He is the recipient of the Distinguished Scientific Award for Early Career Contributions in Psychology from the American Psychological Association (2008), the Troland Research Award from the National Academy of Sciences (2011), the Howard Crosby Warren Medal from the Society of Experimental Psychologists (2016), the R&D Magazine Innovator of the Year award (2018), and a MacArthur Fellowship (2019). He is a fellow of the Cognitive Science Society, the Society for Experimental Psychologists, and a member of the American Academy of Arts and Sciences.

Website: `https://web.mit.edu/cocosci/josh.html`

Session 14A Overview – Wednesday, July 8, 2020 17:00–18:00

Track A *Generation-13* Abstracts	[TACL] A Knowledge-Enhanced Pretraining Model for Commonsense Story Generation *Guan, Huang, Huang, Zhao, and Zhu* [Website][PDF]	BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension *Lewis, Liu, Goyal, Ghazvininejad, Mohamed, Levy, Stoyanov, and Zettlemoyer* [Website][PDF]	BLEURT: Learning Robust Metrics for Text Generation *Sellam, Das, and Parikh* [Website][PDF]	Distilling Knowledge Learned in BERT for Text Generation *Chen, Gan, Cheng, Liu, and Liu* [Website][PDF]	ESPRIT: Explaining Solutions to Physical Reasoning Tasks *Rajani, Zhang, Tan, Zheng, Weiss, Vyas, Gupta, Xiong, Socher, and Radev* [Website][PDF]
	Iterative Edit-Based Unsupervised Sentence Simplification *Kumar, Mou, Golab, and Vechtomova* [Website][PDF]	Logical Natural Language Generation from Open-Domain Tables *Chen, Chen, Su, Chen, and Wang* [Website][PDF]	Neural CRF Model for Sentence Alignment in Text Simplification *Jiang, Maddela, Lan, Zhong, and Xu* [Website][PDF]	One Size Does Not Fit All: Generating and Evaluating Variable Number of Keyphrases *Yuan, Wang, Meng, Thaker, Brusilovsky, He, and Trischler* [Website][PDF]	R^3: Reverse, Retrieve, and Rank for Sarcasm Generation with Commonsense Knowledge *Chakrabarty, Ghosh, Muresan, and Peng* [Website][PDF]
	Shape of Synth to Come: Why We Should Use Synthetic Data for English Surface Realization *Elder, Burke, O'Connor, and Foster* [Website][PDF]	Structural Information Preserving for Graph-to-Text Generation *Song, Wang, Su, Zhang, Xu, Ge, and Yu* [Website][PDF]	[TACL] Syntax-guided Controlled Generation of Paraphrases *Kumar, Ahuja, Vadapalli, and Talukdar* [Website][PDF]		
Track B *Information Extraction-9* Abstracts	A Joint Neural Model for Information Extraction with Global Features *Lin, Ji, Huang, and Wu* [Website][PDF]	A Two-Stage Masked LM Method for Term Set Expansion *Kushilevitz, Markovitch, and Goldberg* [Website][PDF]	Document-Level Event Role Filler Extraction using Multi-Granularity Contextualized Encoding *Du and Cardie* [Website][PDF]	Exploiting the Syntax-Model Consistency for Neural Relation Extraction *Pouran Ben Veyseh, Dernoncourt, Dou, and Nguyen* [Website][PDF]	From English to Code-Switching: Transfer Learning with Strong Morphological Clues *Aguilar and Solorio* [Website][PDF]
	[TACL] Improving Candidate Generation for Low-resource Cross-lingual Entity Linking *Zhou, Rijhwani, Wieting, Carbonell, and Neubig* [Website][PDF]	Learning Interpretable Relationships between Entities, Relations and Concepts via Bayesian Structure Learning on Open Domain Facts *Zhang, Sun, Feng, and Li* [Website][PDF]	Multi-Sentence Argument Linking *Ebner, Xia, Culkin, Rawlins, and Van Durme* [Website][PDF]	Rationalizing Medical Relation Prediction from Corpus-level Statistics *Wang, Lee, Lin, and Sun* [Website][PDF]	Sources of Transfer in Multilingual Named Entity Recognition *Mueller, Andrews, and Dredze* [Website][PDF]
	ZeroShotCeres: Zero-Shot Relation Extraction from Semi-Structured Webpages *Lockard, Shiralkar, Dong, and Hajishirzi* [Website][PDF]	Soft Gazetteers for Low-Resource Named Entity Recognition *Rijhwani, Zhou, Neubig, and Carbonell* [Website][PDF]			

Track C *Information Retrieval and Text Mining-7* Abstracts	A Prioritization Model for Suicidality Risk Assessment *Shing, Resnik, and Oard* [Website][PDF]	CluHTM - Semantic Hierarchical Topic Modeling based on CluWords *Viegas, Cunha, Gomes, Pereira, Rocha, and Goncalves* [Website][PDF]	Document Translation vs. Query Translation for Cross-Lingual Information Retrieval in the Medical Domain *Saleh and Pecina* [Website][PDF]	Empower Entity Set Expansion via Language Model Probing *Zhang, Shen, Shang, and Han* [Website][PDF]	Feature Projection for Improved Text Classification *Qin, Hu, and Liu* [Website][PDF]
Track D *Language Grounding to Vision, Robotics and Beyond-8* Abstracts	A negative case analysis of visual grounding methods for VQA *Shrestha, Kafle, and Kanan* [Website][PDF]	CompGuessWhat?!: A Multi-task Evaluation Framework for Grounded Language Learning *Suglia, Konstas, Vanzo, Bastianelli, Elliott, Frank, and Lemon* [Website][PDF]	History for Visual Dialog: Do we really need it? *Agarwal, Bui, Lee, Konstas, and Rieser* [Website][PDF]	Mapping Natural Language Instructions to Mobile UI Action Sequences *Li, He, Zhou, Zhang, and Baldridge* [Website][PDF]	Multi-agent Communication meets Natural Language: Synergies between Functional and Structural Language Learning *Lazaridou, Potapenko, and Tieleman* [Website][PDF]
	TVQA+: Spatio-Temporal Grounding for Video Question Answering *Lei, Yu, Berg, and Bansal* [Website][PDF]	Unsupervised Multimodal Neural Machine Translation with Pseudo Visual Pivoting *Huang, Hu, Chang, and Hauptmann* [Website][PDF]			
Track E *Phonology, Morphology and Word Segmentation-6* Abstracts	A Multitask Learning Approach for Diacritic Restoration *Alqahtani, Mishra, and Diab* [Website][PDF]	Frugal Paradigm Completion *Erdmann, Kenter, Becker, and Schallhart* [Website][PDF]	Improving Chinese Word Segmentation with Wordhood Memory Networks *Tian, Song, Xia, Zhang, and Wang* [Website][PDF]	Joint Chinese Word Segmentation and Part-of-speech Tagging via Two-way Attentions of Auto-analyzed Knowledge *Tian, Song, Ao, Xia, Quan, Zhang, and Wang* [Website][PDF]	Joint Diacritization, Lemmatization, Normalization, and Fine-Grained Morphological Tagging *Zalmout and Habash* [Website][PDF]
	Phonetic and Visual Priors for Decipherment of Informal Romanization *Ryskina, Gormley, and Berg-Kirkpatrick* [Website][PDF]	[TACL] Phonotactic Complexity and its Trade-offs *Pimentel, Roark, and Cotterell* [Website][PDF]	The Paradigm Discovery Problem *Erdmann, Elsner, Wu, Cotterell, and Habash* [Website][PDF]	Supervised Grapheme-to-Phoneme Conversion of Orthographic Schwas in Hindi and Punjabi *Arora, Gessler, and Schneider* [Website][PDF]	
Track F *Sentence Level-9* Abstracts	Active Learning for Coreference Resolution using Discrete Annotation *Li, Stanovsky, and Zettlemoyer* [Website][PDF]	Beyond Possession Existence: Duration and Co-Possession *Chinnappa, Murugan, and Blanco* [Website][PDF]	[TACL] Decoding Brain Activity Associated with Literal and Metaphoric Sentence Comprehension using Distributional Semantic Models *Djokic, Maillard, Bulat, and Shutova* [Website][PDF]	Don't Stop Pretraining: Adapt Language Models to Domains and Tasks *Gururangan, Marasović, Swayamdipta, Lo, Beltagy, Downey, and Smith* [Website][PDF]	Estimating Mutual Information Between Dense Word Embeddings *Zhelezniak, Savkov, and Hammerla* [Website][PDF]

	Exploring Unexplored Generalization Challenges for Cross-Database Semantic Parsing *Suhr, Chang, Shaw, and Lee* [Website][PDF]	Predicting the Focus of Negation: Model and Error Analysis *Hossain, Hamilton, Palmer, and Blanco* [Website][PDF]	RAT-SQL: Relation-Aware Schema Encoding and Linking for Text-to-SQL Parsers *Wang, Shin, Liu, Polozov, and Richardson* [Website][PDF]	Structured Tuning for Semantic Role Labeling *Li, Jawale, Palmer, and Srikumar* [Website][PDF]	TaBERT: Pretraining for Joint Understanding of Textual and Tabular Data *Yin, Neubig, Yih, and Riedel* [Website][PDF]
	Universal Decompositional Semantic Parsing *Stengel-Eskin, White, Zhang, and Van Durme* [Website][PDF]	Unsupervised Cross-lingual Representation Learning at Scale *Conneau, Khandelwal, Goyal, Chaudhary, Wenzek, Guzmán, Grave, Ott, Zettlemoyer, and Stoyanov* [Website][PDF]			
Track G *Student Research Workshop* Abstracts	Topic Balancing with Additive Regularization of Topic Models *Veselova and Vorontsov* [Website][PDF]	Combining Subword Representations into Word-level Representations in the Transformer Architecture *Casas, Costa-jussà, and Fonollosa* [Website][PDF]	Exploring Interpretability in Event Extraction: Multitask Learning of a Neural Event Classifier and an Explanation Decoder *Tang, Hahn-Powell, and Surdeanu* [Website][PDF]	Dominance as an Indicator of Rapport and Learning in Human-Agent Communication *Buddemeyer, Tian, and Walker* [Website]	

Session 14A Details

Session 14A: Generation-13

[TACL] A Knowledge-Enhanced Pretraining Model for Commonsense Story Generation [Website] [PDF]
Jian Guan, Fei Huang, Minlie Huang, Zhihao Zhao, and Xiaoyan Zhu 17:00–18:00

Story generation, namely generating a reasonable story from a leading context, is an important but challenging task. In spite of the success in modeling fluency and local coherence, existing neural language generation models (e.g., GPT-2) still suffer from repetition, logic conflicts, and lack of long-range coherence in generated stories. We conjecture that this is because of the difficulty of associating relevant commonsense knowledge, understanding the causal relationships, and planning entities and events with proper temporal order. In this paper, we devise a knowledge-enhanced pretraining model for commonsense story generation. We propose to utilize commonsense knowledge from external knowledge bases to generate reasonable stories. To further capture the causal and temporal dependencies between the sentences in a reasonable story, we employ multi-task learning which combines a discriminative objective to distinguish true and fake stories during fine-tuning. Automatic and manual evaluation shows that our model can generate more reasonable stories than state-of-the-art baselines, particularly in terms of logic and global coherence.

BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension [Website] [PDF]
Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer 17:00–18:00

We present BART, a denoising autoencoder for pretraining sequence-to-sequence models. BART is trained by (1) corrupting text with an arbitrary noising function, and (2) learning a model to reconstruct the original text. It uses a standard Tranformer-based neural machine translation architecture which, despite its simplicity, can be seen as generalizing BERT (due to the bidirectional encoder), GPT (with the left-to-right decoder), and other recent pretraining schemes. We evaluate a number of noising approaches, finding the best performance by both randomly shuffling the order of sentences and using a novel in-filling scheme, where spans of text are replaced with a single mask token. BART is particularly effective when fine tuned for text generation but also works well for comprehension tasks. It matches the performance of RoBERTa on GLUE and SQuAD, and achieves new state-of-the-art results on a range of abstractive dialogue, question answering, and summarization tasks, with gains of up to 3.5 ROUGE. BART also provides a 1.1 BLEU increase over a back-translation system for machine translation, with only target language pretraining. We also replicate other pretraining schemes within the BART framework, to understand their effect on end-task performance.

BLEURT: Learning Robust Metrics for Text Generation [Website] [PDF]
Thibault Sellam, Dipanjan Das, and Ankur Parikh 17:00–18:00

Text generation has made significant advances in the last few years. Yet, evaluation metrics have lagged behind, as the most popular choices (e.g., BLEU and ROUGE) may correlate poorly with human judgment. We propose BLEURT, a learned evaluation metric for English based on BERT. BLEURT can model human judgment with a few thousand possibly biased training examples. A key aspect of our approach is a novel pre-training scheme that uses millions of synthetic examples to help the model generalize. BLEURT provides state-of-the-art results on the last three years of the WMT Metrics shared task and the WebNLG data set. In contrast to a vanilla BERT-based approach, it yields superior results even when the training data is scarce and out-of-distribution.

Distilling Knowledge Learned in BERT for Text Generation [Website] [PDF]
Yen-Chun Chen, Zhe Gan, Yu Cheng, Jingzhou Liu, and Jingjing Liu 17:00–18:00

Large-scale pre-trained language model such as BERT has achieved great success in language understanding tasks. However, it remains an open question how to utilize BERT for language generation. In this paper, we present a novel approach, Conditional Masked Language Modeling (C-MLM), to enable the finetuning of BERT on target generation tasks. The finetuned BERT (teacher) is exploited as extra supervision to improve conventional Seq2Seq models (student) for better text generation performance. By leveraging BERT's idiosyncratic bidirectional nature, distilling knowledge learned in BERT can encourage auto-regressive Seq2Seq models to plan ahead, imposing global sequence-level supervision for coherent text generation. Experiments show that the proposed approach significantly outperforms strong Transformer baselines on multiple language generation tasks such as machine translation and text summarization. Our proposed model also achieves new state of the art on IWSLT German-English and English-Vietnamese MT datasets.

ESPRIT: Explaining Solutions to Physical Reasoning Tasks [Website] [PDF]
Nazneen Fatema Rajani, Rui Zhang, Yi Chern Tan, Stephan Zheng, Jeremy Weiss, Aadit Vyas, Abhijit Gupta, Caiming Xiong, Richard Socher, and Dragomir Radev 17:00–18:00

Neural networks lack the ability to reason about qualitative physics and so cannot generalize to scenarios and tasks unseen during training. We propose ESPRIT, a framework for commonsense reasoning about qualitative physics in natural language that generates interpretable descriptions of physical events. We use a two-step approach of first identifying the pivotal physical events in an environment and then generating natural language descriptions of those events using a data-to-text approach. Our framework learns to generate explanations of how the physical simulation will causally evolve so that an agent or a human can easily reason about a solution using those interpretable

descriptions. Human evaluations indicate that ESPRIT produces crucial fine-grained details and has high coverage of physical concepts compared to even human annotations. Dataset, code and documentation are available at https://github.com/salesforce/esprit.

Iterative Edit-Based Unsupervised Sentence Simplification [Website][PDF]
Dhruv Kumar, Lili Mou, Lukasz Golab, and Olga Vechtomova 17:00–18:00

We present a novel iterative, edit-based approach to unsupervised sentence simplification. Our model is guided by a scoring function involving fluency, simplicity, and meaning preservation. Then, we iteratively perform word and phrase-level edits on the complex sentence. Compared with previous approaches, our model does not require a parallel training set, but is more controllable and interpretable. Experiments on Newsela and WikiLarge datasets show that our approach is nearly as effective as state-of-the-art supervised approaches.

Logical Natural Language Generation from Open-Domain Tables [Website][PDF]
Wenhu Chen, Jianshu Chen, Yu Su, Zhiyu Chen, and William Yang Wang 17:00–18:00

Neural natural language generation (NLG) models have recently shown remarkable progress in fluency and coherence. However, existing studies on neural NLG are primarily focused on surface-level realizations with limited emphasis on logical inference, an important aspect of human thinking and language. In this paper, we suggest a new NLG task where a model is tasked with generating natural language statements that can be *logically entailed* by the facts in an open-domain semi-structured table. To facilitate the study of the proposed logical NLG problem, we use the existing TabFact dataset~[**chen2019tabfact**] featured with a wide range of logical/symbolic inferences as our testbed, and propose new automatic metrics to evaluate the fidelity of generation models w.r.t. logical inference. The new task poses challenges to the existing monotonic generation frameworks due to the mismatch between sequence order and logical order. In our experiments, we comprehensively survey different generation architectures (LSTM, Transformer, Pre-Trained LM) trained with different algorithms (RL, Adversarial Training, Coarse-to-Fine) on the dataset and made following observations: 1) Pre-Trained LM can significantly boost both the fluency and logical fidelity metrics, 2) RL and Adversarial Training are trading fluency for fidelity, 3) Coarse-to-Fine generation can help partially alleviate the fidelity issue while maintaining high language fluency. The code and data are available at https://github.com/wenhuchen/LogicNLG.

Neural CRF Model for Sentence Alignment in Text Simplification [Website][PDF]
Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu 17:00–18:00

The success of a text simplification system heavily depends on the quality and quantity of complex-simple sentence pairs in the training corpus, which are extracted by aligning sentences between parallel articles. To evaluate and improve sentence alignment quality, we create two manually annotated sentence-aligned datasets from two commonly used text simplification corpora, Newsela and Wikipedia. We propose a novel neural CRF alignment model which not only leverages the sequential nature of sentences in parallel documents but also utilizes a neural sentence pair model to capture semantic similarity. Experiments demonstrate that our proposed approach outperforms all the previous work on monolingual sentence alignment task by more than 5 points in F1. We apply our CRF aligner to construct two new text simplification datasets, NEWSELA-AUTO and WIKI-AUTO, which are much larger and of better quality compared to the existing datasets. A Transformer-based seq2seq model trained on our datasets establishes a new state-of-the-art for text simplification in both automatic and human evaluation.

One Size Does Not Fit All: Generating and Evaluating Variable Number of Keyphrases [Website][PDF]
Xingdi Yuan, Tong Wang, Rui Meng, Khushboo Thaker, Peter Brusilovsky, Daqing He, and Adam Trischler 17:00–18:00

Different texts shall by nature correspond to different number of keyphrases. This desideratum is largely missing from existing neural keyphrase generation models. In this study, we address this problem from both modeling and evaluation perspectives. We first propose a recurrent generative model that generates multiple keyphrases as delimiter-separated sequences. Generation diversity is further enhanced with two novel techniques by manipulating decoder hidden states. In contrast to previous approaches, our model is capable of generating diverse keyphrases and controlling number of outputs. We further propose two evaluation metrics tailored towards the variable-number generation. We also introduce a new dataset StackEx that expands beyond the only existing genre (i.e., academic writing) in keyphrase generation tasks. With both previous and new evaluation metrics, our model outperforms strong baselines on all datasets.

R^3: Reverse, Retrieve, and Rank for Sarcasm Generation with Commonsense Knowledge [Website][PDF]
Tuhin Chakrabarty, Debanjan Ghosh, Smaranda Muresan, and Nanyun Peng 17:00–18:00

We propose an unsupervised approach for sarcasm generation based on a non-sarcastic input sentence. Our method employs a retrieve-and-edit framework to instantiate two major characteristics of sarcasm: reversal of valence and semantic incongruity with the context, which could include shared commonsense or world knowledge between the speaker and the listener. While prior works on sarcasm generation predominantly focus on context incongruity, we show that combining valence reversal and semantic incongruity based on the commonsense knowledge generates sarcasm of higher quality. Human evaluation shows that our system generates sarcasm better than humans 34% of the time, and better than a reinforced hybrid baseline 90% of the time.

Shape of Synth to Come: Why We Should Use Synthetic Data for English Surface Realization [Website][PDF]
Henry Elder, Robert Burke, Alexander O'Connor, and Jennifer Foster 17:00–18:00

The Surface Realization Shared Tasks of 2018 and 2019 were Natural Language Generation shared tasks with the goal of exploring approaches to surface realization from Universal-Dependency-like trees to surface strings for several languages. In the 2018 shared task there was very little difference in the absolute performance of systems trained with and without additional, synthetically created data, and a new rule prohibiting the use of synthetic data was introduced for the 2019 shared task. Contrary to the findings of the 2018 shared task, we show, in experiments on the English 2018 dataset, that the use of synthetic data can have a substantial positive effect – an improvement of almost 8 BLEU points for a previously state-of-the-art system. We analyse the effects of synthetic data, and we argue that its use should be encouraged rather than prohibited so that future research efforts continue to explore systems that can take advantage of such data.

Structural Information Preserving for Graph-to-Text Generation [Website][PDF]
Linfeng Song, Ante Wang, Jinsong Su, Yue Zhang, Kun Xu, Yubin Ge, and Dong Yu 17:00–18:00

The task of graph-to-text generation aims at producing sentences that preserve the meaning of input graphs. As a crucial defect, the current state-of-the-art models may mess up or even drop the core structural information of input graphs when generating outputs. We propose to tackle this problem by leveraging richer training signals that can guide our model for preserving input information. In particular, we introduce two types of autoencoding losses, each individually focusing on different aspects (a.k.a. views) of input graphs. The losses are then back-propagated to better calibrate our model via multi-task training. Experiments on two benchmarks for graph-to-text generation show the effectiveness of our approach over a state-of-the-art baseline.

[TACL] Syntax-guided Controlled Generation of Paraphrases [Website][PDF]
Ashutosh Kumar, Kabir Ahuja, Raghuram Vadapalli, and Partha Talukdar 17:00–18:00

Given a sentence (e.g., "I like mangoes") and a constraint (e.g., negative sentiment), the goal of controlled text generation is to produce a sentence that adapts the input sentence to meet the requirements of the constraint (e.g., "I hate mangoes"). Going beyond such simple constraints, recent works have started exploring the incorporation of complex syntactic-guidance as constraints in the task of controlled paraphrase generation. In these methods, syntactic-guidance is sourced from a separate exemplar sentence. However, these prior works have only utilized limited syntactic information available in the parse tree of the exemplar sentence. We address this limitation in the paper and propose Syntax Guided Controlled Paraphraser (SGCP), an end-to-end framework for syntactic paraphrase generation. We find that S GCP can generate syntax-conforming sentences while not compromising on relevance. We perform extensive automated and human evaluations over multiple real-world datasets to demonstrate the efficacy of SGCP over state-of-the-art baselines. To drive future research, we have made SGCP's source code available.

Session 14A: Information Extraction-9

A Joint Neural Model for Information Extraction with Global Features [Website][PDF]
Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu 17:00–18:00

Most existing joint neural models for Information Extraction (IE) use local task-specific classifiers to predict labels for individual instances (e.g., trigger, relation) regardless of their interactions. For example, a victim of a die event is likely to be a victim of an attack event in the same sentence. In order to capture such cross-subtask and cross-instance inter-dependencies, we propose a joint neural framework, OneIE, that aims to extract the globally optimal IE result as a graph from an input sentence. OneIE performs end-to-end IE in four stages: (1) Encoding a given sentence as contextualized word representations; (2) Identifying entity mentions and event triggers as nodes; (3) Computing label scores for all nodes and their pairwise links using local classifiers; (4) Searching for the globally optimal graph with a beam decoder. At the decoding stage, we incorporate global features to capture the cross-subtask and cross-instance interactions. Experiments show that adding global features improves the performance of our model and achieves new state of-the-art on all subtasks. In addition, as OneIE does not use any language-specific feature, we prove it can be easily applied to new languages or trained in a multilingual manner.

A Two-Stage Masked LM Method for Term Set Expansion [Website][PDF]
Guy Kushilevitz, Shaul Markovitch, and Yoav Goldberg 17:00–18:00

We tackle the task of Term Set Expansion (TSE): given a small seed set of example terms from a semantic class, finding more members of that class. The task is of great practical utility, and also of theoretical utility as it requires generalization from few examples. Previous approaches to the TSE task can be characterized as either distributional or pattern-based. We harness the power of neural masked language models (MLM) and propose a novel TSE algorithm, which combines the pattern-based and distributional approaches. Due to the small size of the seed set, fine-tuning methods are not effective, calling for more creative use of the MLM. The gist of the idea is to use the MLM to first mine for informative patterns with respect to the seed set, and then to obtain more members of the seed class by generalizing these patterns. Our method outperforms state-of-the-art TSE algorithms. Implementation is available at: https://github.com/ guykush/TermSetExpansion-MPB/

Document-Level Event Role Filler Extraction using Multi-Granularity Contextualized Encoding [Website][PDF]
Xinya Du and Claire Cardie 17:00–18:00

Few works in the literature of event extraction have gone beyond individual sentences to make extraction decisions. This is problematic when the information needed to recognize an event argument is spread across multiple sentences. We argue that document-level event extraction is a difficult task since it requires a view of a larger context to determine which spans of text correspond to event role fillers. We first investigate how end-to-end neural sequence models (with pre-trained language model representations) perform on document-level role filler extraction, as well as how the length of context captured affects the models' performance. To dynamically aggregate information captured by neural representations learned at different levels of granularity (e.g., the sentence- and paragraph-level), we propose a novel multi-granularity reader. We evaluate our models on the MUC-4 event extraction dataset, and show that our best system performs substantially better than prior work. We also report findings on the relationship between context length and neural model performance on the task.

Exploiting the Syntax-Model Consistency for Neural Relation Extraction [Website][PDF]
Amir Pouran Ben Veyseh, Franck Dernoncourt, Dejing Dou, and Thien Huu Nguyen 17:00–18:00

This paper studies the task of Relation Extraction (RE) that aims to identify the semantic relations between two entity mentions in text. In the deep learning models for RE, it has been beneficial to incorporate the syntactic structures from the dependency trees of the input sentences. In such models, the dependency trees are often used to directly structure the network architectures or to obtain the dependency relations between the word pairs to inject the syntactic information into the models via multi-task learning. The major problem with these approaches is the lack of generalization beyond the syntactic structures in the training data or the failure to capture the syntactic importance of the words for RE. In order to overcome these issues, we propose a novel deep learning model for RE that uses the dependency trees to extract the syntax-based importance scores for the words, serving as a tree representation to introduce syntactic information into the models with greater generalization. In particular, we leverage Ordered-Neuron Long-Short Term Memory Networks (ON-LSTM) to infer the model-based importance scores for RE for every word in the sentences that are then regulated to be consistent with the syntax-based scores to enable syntactic information injection. We perform extensive experiments to demonstrate the effectiveness of the proposed method, leading to the state-of-the-art performance on three RE benchmark datasets.

From English to Code-Switching: Transfer Learning with Strong Morphological Clues [Website][PDF]
Gustavo Aguilar and Thamar Solorio 17:00–18:00

Linguistic Code-switching (CS) is still an understudied phenomenon in natural language processing. The NLP community has mostly focused on monolingual and multi-lingual scenarios, but little attention has been given to CS in particular. This is partly because of the lack of resources and annotated data, despite its increasing occurrence in social media platforms. In this paper, we aim at adapting monolingual models to code-switched text in various tasks. Specifically, we transfer English knowledge from a pre-trained ELMo model to different code-switched language pairs (i.e., Nepali-English, Spanish-English, and Hindi-English) using the task of language identification. Our method, CS-ELMo, is an extension of ELMo with a simple yet effective position-aware attention mechanism inside its character convolutions. We show the effectiveness of this transfer learning step by outperforming multilingual BERT and ho-

mologous CS-unaware ELMo models and establishing a new state of the art in CS tasks, such as NER and POS tagging. Our technique can be expanded to more English-paired code-switched languages, providing more resources to the CS community.

[TACL] Improving Candidate Generation for Low-resource Cross-lingual Entity Linking [Website][PDF]
Shuyan Zhou, Shruti Rijhwani, John Wieting, Jaime Carbonell, and Graham Neubig 17:00–18:00

Cross-lingual entity linking (XEL) is the task of finding referents in a target-language knowledge base (KB) for mentions extracted from source-language texts. The first step of (X)EL is candidate generation, which retrieves a list of plausible candidate entities from the target-language KB for each mention. Approaches based on resources from Wikipedia have proven successful in the realm of relatively high-resource languages (HRL), but these do not extend well to low-resource languages (LRL) with few, if any, Wikipedia pages. Recently, transfer learning methods have been shown to reduce the demand for resources in the LRL by utilizing resources in closely-related languages, but the performance still lags far behind their high-resource counterparts. In this paper, we first assess the problems faced by current entity candidate generation methods for low-resource XEL, then propose three improvements that (1) reduce the disconnect between entity mentions and KB entries, and (2) improve the robustness of the model to low-resource scenarios. The methods are simple, but effective: we experiment with our approach on seven XEL datasets and find that they yield an average gain of 16.9% in Top-30 gold candidate recall, compared to state-of-the-art baselines. Our improved model also yields an average gain of 7.9% in in-KB accuracy of end-to-end XEL.

Learning Interpretable Relationships between Entities, Relations and Concepts via Bayesian Structure Learning on Open Domain Facts [Website][PDF]
Jingyuan Zhang, Mingming Sun, Yue Feng, and Ping Li 17:00–18:00

Concept graphs are created as universal taxonomies for text understanding in the open-domain knowledge. The nodes in concept graphs include both entities and concepts. The edges are from entities to concepts, showing that an entity is an instance of a concept. In this paper, we propose the task of learning interpretable relationships from open-domain facts to enrich and refine concept graphs. The Bayesian network structures are learned from open-domain facts as the interpretable relationships between relations of facts and concepts of entities. We conduct extensive experiments on public English and Chinese datasets. Compared to the state-of-the-art methods, the learned network structures help improving the identification of concepts for entities based on the relations of entities on both datasets.

Multi-Sentence Argument Linking [Website][PDF]
Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, and Benjamin Van Durme 17:00–18:00

We present a novel document-level model for finding argument spans that fill an event's roles, connecting related ideas in sentence-level semantic role labeling and coreference resolution. Because existing datasets for cross-sentence linking are small, development of our neural model is supported through the creation of a new resource, Roles Across Multiple Sentences (RAMS), which contains 9,124 annotated events across 139 types. We demonstrate strong performance of our model on RAMS and other event-related datasets.

Rationalizing Medical Relation Prediction from Corpus-level Statistics [Website][PDF]
Zhen Wang, Jennifer Lee, Simon Lin, and Huan Sun 17:00–18:00

Nowadays, the interpretability of machine learning models is becoming increasingly important, especially in the medical domain. Aiming to shed some light on how to rationalize medical relation prediction, we present a new interpretable framework inspired by existing theories on how human memory works, e.g., theories of recall and recognition. Given the corpus-level statistics, i.e., a global co-occurrence graph of a clinical text corpus, to predict the relations between two entities, we first recall rich contexts associated with the target entities, and then recognize relational interactions between these contexts to form model rationales, which will contribute to the final prediction. We conduct experiments on a real-world public clinical dataset and show that our framework can not only achieve competitive predictive performance against a comprehensive list of neural baseline models, but also present rationales to justify its prediction. We further collaborate with medical experts deeply to verify the usefulness of our model rationales for clinical decision making.

Sources of Transfer in Multilingual Named Entity Recognition [Website][PDF]
David Mueller, Nicholas Andrews, and Mark Dredze 17:00–18:00

Named-entities are inherently multilingual, and annotations in any given language may be limited. This motivates us to consider *polyglot* named-entity recognition (NER), where one model is trained using annotated data drawn from more than one language. However, a straightforward implementation of this simple idea does not always work in practice: naive training of NER models using annotated data drawn from multiple languages consistently underperforms models trained on monolingual data alone, despite having access to more training data. The starting point of this paper is a simple solution to this problem, in which polyglot models are *fine-tuned* on monolingual data to consistently and significantly outperform their monolingual counterparts. To explain this phenomena, we explore the sources of multilingual transfer in polyglot NER models and examine the weight structure of polyglot models compared to their monolingual counterparts. We find that polyglot models efficiently share many parameters across languages and that fine-tuning may utilize a large number of those parameters.

ZeroShotCeres: Zero-Shot Relation Extraction from Semi-Structured Webpages [Website][PDF]
Colin Lockard, Prashant Shiralkar, Xin Luna Dong, and Hannaneh Hajishirzi 17:00–18:00

In many documents, such as semi-structured webpages, textual semantics are augmented with additional information conveyed using visual elements including layout, font size, and color. Prior work on information extraction from semi-structured websites has required learning an extraction model specific to a given template via either manually

labeled or distantly supervised data from that template. In this work, we propose a solution for "zero-shot" open-domain relation extraction from webpages with a previously unseen template, including from websites with little overlap with existing sources of knowledge for distant supervision and websites in entirely new subject verticals. Our model uses a graph neural network-based approach to build a rich representation of text fields on a webpage and the relationships between them, enabling generalization to new templates. Experiments show this approach provides a 31% F1 gain over a baseline for zero-shot extraction in a new subject vertical.

Soft Gazetteers for Low-Resource Named Entity Recognition [Website] [PDF]
Shruti Rijhwani, Shuyan Zhou, Graham Neubig, and Jaime Carbonell 17:00–18:00

Traditional named entity recognition models use gazetteers (lists of entities) as features to improve performance. Although modern neural network models do not require such hand-crafted features for strong performance, recent work has demonstrated their utility for named entity recognition on English data. However, designing such features for low-resource languages is challenging, because exhaustive entity gazetteers do not exist in these languages. To address this problem, we propose a method of "soft gazetteers" that incorporates ubiquitously available information from English knowledge bases, such as Wikipedia, into neural named entity recognition models through cross-lingual entity linking. Our experiments on four low-resource languages show an average improvement of 4 points in F1 score.

Session 14A: Information Retrieval and Text Mining-7

A Prioritization Model for Suicidality Risk Assessment [Website][PDF]
Han-Chin Shing, Philip Resnik, and Douglas Oard 17:00–18:00

We reframe suicide risk assessment from social media as a ranking problem whose goal is maximizing detection of severely at-risk individuals given the time available. Building on measures developed for resource-bounded document retrieval, we introduce a well founded evaluation paradigm, and demonstrate using an expert-annotated test collection that meaningful improvements over plausible cascade model baselines can be achieved using an approach that jointly ranks individuals and their social media posts.

CluHTM - Semantic Hierarchical Topic Modeling based on CluWords [Website][PDF]
Felipe Viegas, Washington Cunha, Christian Gomes, Antônio Pereira, Leonardo Rocha, and Marcos Goncalves 17:00–18:00

Hierarchical Topic modeling (HTM) exploits latent topics and relationships among them as a powerful tool for data analysis and exploration. Despite advantages over traditional topic modeling, HTM poses its own challenges, such as (1) topic incoherence, (2) unreasonable (hierarchical) structure, and (3) issues related to the definition of the "ideal" number of topics and depth of the hierarchy. In this paper, we advance the state-of-the-art on HTM by means of the design and evaluation of CluHTM, a novel non-probabilistic hierarchical matrix factorization aimed at solving the specific issues of HTM. CluHTM's novel contributions include: (i) the exploration of richer text representation that encapsulates both, global (dataset level) and local semantic information – when combined, these pieces of information help to solve the topic incoherence problem as well as issues related to the unreasonable structure; (ii) the exploitation of a stability analysis metric for defining the number of topics and the "shape" the hierarchical structure. In our evaluation, considering twelve datasets and seven state-of-the-art baselines, CluHTM outperformed the baselines in the vast majority of the cases, with gains of around \$500%\$ over the strongest state-of-the-art baselines. We also provide qualitative and quantitative statistical analyses of why our solution works so well.

Document Translation vs. Query Translation for Cross-Lingual Information Retrieval in the Medical Domain [Website][PDF]
Shadi Saleh and Pavel Pecina 17:00–18:00

We present a thorough comparison of two principal approaches to Cross-Lingual Information Retrieval: document translation (DT) and query translation (QT). Our experiments are conducted using the cross-lingual test collection produced within the CLEF eHealth information retrieval tasks in 2013—2015 containing English documents and queries in several European languages. We exploit the Statistical Machine Translation (SMT) and Neural Machine Translation (NMT) paradigms and train several domain-specific and task-specific machine translation systems to translate the non-English queries into English (for the QT approach) and the English documents to all the query languages (for the DT approach). The results show that the quality of QT by SMT is sufficient enough to outperform the retrieval results of the DT approach for all the languages. NMT then further boosts translation quality and retrieval quality for both QT and DT for most languages, but still, QT provides generally better retrieval results than DT.

Empower Entity Set Expansion via Language Model Probing [Website][PDF]
Yunyi Zhang, Jiaming Shen, Jingbo Shang, and Jiawei Han 17:00–18:00

Entity set expansion, aiming at expanding a small seed entity set with new entities belonging to the same semantic class, is a critical task that benefits many downstream NLP and IR applications, such as question answering, query understanding, and taxonomy construction. Existing set expansion methods bootstrap the seed entity set by adaptively selecting context features and extracting new entities. A key challenge for entity set expansion is to avoid selecting ambiguous context features which will shift the class semantics and lead to accumulative errors in later iterations. In this study, we propose a novel iterative set expansion framework that leverages automatically generated class names to address the semantic drift issue. In each iteration, we select one positive and several negative class names by probing a pre-trained language model, and further score each candidate entity based on selected class names. Experiments on two datasets show that our framework generates high-quality class names and outperforms previous state-of-the-art methods significantly.

Feature Projection for Improved Text Classification [Website][PDF]
Qi Qin, Wenpeng Hu, and Bing Liu 17:00–18:00

In classification, there are usually some good features that are indicative of class labels. For example, in sentiment classification, words like good and nice are indicative of the positive sentiment and words like bad and terrible are indicative of the negative sentiment. However, there are also many common features (e.g., words) that are not indicative of any specific class (e.g., voice and screen, which are common to both sentiment classes and are not discriminative for classification). Although deep learning has made significant progresses in generating discriminative features through its powerful representation learning, we believe there is still room for improvement. In this paper, we propose a novel angle to further improve this representation learning, i.e., feature projection. This method projects existing features into the orthogonal space of the common features. The resulting projection is thus perpendicular to the common features and more discriminative for classification. We apply this new method to improve CNN, RNN, Transformer, and Bert based text classification and obtain markedly better results.

Session 14A: Language Grounding to Vision, Robotics and Beyond-8

A negative case analysis of visual grounding methods for VQA [Website][PDF]
Robik Shrestha, Kushal Kafle, and Christopher Kanan 17:00–18:00

Existing Visual Question Answering (VQA) methods tend to exploit dataset biases and spurious statistical correlations, instead of producing right answers for the right reasons. To address this issue, recent bias mitigation methods for VQA propose to incorporate visual cues (e.g., human attention maps) to better ground the VQA models, showcasing impressive gains. However, we show that the performance improvements are not a result of improved visual grounding, but a regularization effect which prevents over-fitting to linguistic priors. For instance, we find that it is not actually necessary to provide proper, human-based cues; random, insensible cues also result in similar improvements. Based on this observation, we propose a simpler regularization scheme that does not require any external annotations and yet achieves near state-of-the-art performance on VQA-CPv2.

CompGuessWhat?!: A Multi-task Evaluation Framework for Grounded Language Learning [Website][PDF]
Alessandro Suglia, Ioannis Konstas, Andrea Vanzo, Emanuele Bastianelli, Desmond Elliott, Stella Frank, and Oliver Lemon 17:00–18:00

Approaches to Grounded Language Learning are commonly focused on a single task-based final performance measure which may not depend on desirable properties of the learned hidden representations, such as their ability to predict object attributes or generalize to unseen situations. To remedy this, we present GroLLA, an evaluation framework for Grounded Language Learning with Attributes based on three sub-tasks: 1) Goal-oriented evaluation; 2) Object attribute prediction evaluation; and 3) Zero-shot evaluation. We also propose a new dataset CompGuessWhat?! as an instance of this framework for evaluating the quality of learned neural representations, in particular with respect to attribute grounding. To this end, we extend the original GuessWhat?! dataset by including a semantic layer on top of the perceptual one. Specifically, we enrich the VisualGenome scene graphs associated with the GuessWhat?! images with several attributes from resources such as VISA and ImSitu. We then compare several hidden state representations from current state-of-the-art approaches to Grounded Language Learning. By using diagnostic classifiers, we show that current models' learned representations are not expressive enough to encode object attributes (average F1 of 44.27). In addition, they do not learn strategies nor representations that are robust enough to perform well when novel scenes or objects are involved in gameplay (zero-shot best accuracy 50.06%).

History for Visual Dialog: Do we really need it? [Website][PDF]
Shubham Agarwal, Trung Bui, Joon-Young Lee, Ioannis Konstas, and Verena Rieser 17:00–18:00

Visual Dialogue involves "understanding" the dialogue history (what has been discussed previously) and the current question (what is asked), in addition to grounding information in the image, to accurately generate the correct response. In this paper, we show that co-attention models which explicitly encode dialoh history outperform models that don't, achieving state-of-the-art performance (72 % NDCG on val set). However, we also expose shortcomings of the crowdsourcing dataset collection procedure, by showing that dialogue history is indeed only required for a small amount of the data, and that the current evaluation metric encourages generic replies. To that end, we propose a challenging subset (VisdialConv) of the VisdialVal set and the benchmark NDCG of 63%.

Mapping Natural Language Instructions to Mobile UI Action Sequences [Website][PDF]
Yang Li, Jiacong He, Xin Zhou, Yuan Zhang, and Jason Baldridge 17:00–18:00

We present a new problem: grounding natural language instructions to mobile user interface actions, and create three new datasets for it. For full task evaluation, we create PixelHelp, a corpus that pairs English instructions with actions performed by people on a mobile UI emulator. To scale training, we decouple the language and action data by (a) annotating action phrase spans in How-To instructions and (b) synthesizing grounded descriptions of actions for mobile user interfaces. We use a Transformer to extract action phrase tuples from long-range natural language instructions. A grounding Transformer then contextually represents UI objects using both their content and screen position and connects them to object descriptions. Given a starting screen and instruction, our model achieves 70.59% accuracy on predicting complete ground-truth action sequences in PixelHelp.

Multi-agent Communication meets Natural Language: Synergies between Functional and Structural Language Learning [Website][PDF]
Angeliki Lazaridou, Anna Potapenko, and Olivier Tieleman 17:00–18:00

We present a method for combining multi-agent communication and traditional data-driven approaches to natural language learning, with an end goal of teaching agents to communicate with humans in natural language. Our starting point is a language model that has been trained on generic, not task-specific language data. We then place this model in a multi-agent self-play environment that generates task-specific rewards used to adapt or modulate the model, turning it into a task-conditional language model. We introduce a new way for combining the two types of learning based on the idea of reranking language model samples, and show that this method outperforms others in communicating with humans in a visual referential communication task. Finally, we present a taxonomy of different types of language drift that can occur alongside a set of measures to detect them.

TVQA+: Spatio-Temporal Grounding for Video Question Answering [Website][PDF]
Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal 17:00–18:00

We present the task of Spatio-Temporal Video Question Answering, which requires intelligent systems to simultaneously retrieve relevant moments and detect referenced visual concepts (people and objects) to answer natural lan-

guage questions about videos. We first augment the TVQA dataset with 310.8K bounding boxes, linking depicted objects to visual concepts in questions and answers. We name this augmented version as TVQA+. We then propose Spatio-Temporal Answerer with Grounded Evidence (STAGE), a unified framework that grounds evidence in both spatial and temporal domains to answer questions about videos. Comprehensive experiments and analyses demonstrate the effectiveness of our framework and how the rich annotations in our TVQA+ dataset can contribute to the question answering task. Moreover, by performing this joint task, our model is able to produce insightful and interpretable spatio-temporal attention visualizations.

Unsupervised Multimodal Neural Machine Translation with Pseudo Visual Pivoting [Website] [PDF]
Po-Yao Huang, Junjie Hu, Xiaojun Chang, and Alexander Hauptmann 17:00–18:00

Unsupervised machine translation (MT) has recently achieved impressive results with monolingual corpora only. However, it is still challenging to associate source-target sentences in the latent space. As people speak different languages biologically share similar visual systems, the potential of achieving better alignment through visual content is promising yet under-explored in unsupervised multimodal MT (MMT). In this paper, we investigate how to utilize visual content for disambiguation and promoting latent space alignment in unsupervised MMT. Our model employs multimodal back-translation and features pseudo visual pivoting in which we learn a shared multilingual visual-semantic embedding space and incorporate visually-pivoted captioning as additional weak supervision. The experimental results on the widely used Multi30K dataset show that the proposed model significantly improves over the state-of-the-art methods and generalizes well when images are not available at the testing time.

Session 14A: Phonology, Morphology and Word Segmentation-6

A Multitask Learning Approach for Diacritic Restoration [Website][PDF]
Sawsan Alqahtani, Ajay Mishra, and Mona Diab 17:00–18:00

In many languages like Arabic, diacritics are used to specify pronunciations as well as meanings. Such diacritics are often omitted in written text, increasing the number of possible pronunciations and meanings for a word. This results in a more ambiguous text making computational processing on such text more difficult. Diacritic restoration is the task of restoring missing diacritics in the written text. Most state-of-the-art diacritic restoration models are built on character level information which helps generalize the model to unseen data, but presumably lose useful information at the word level. Thus, to compensate for this loss, we investigate the use of multi-task learning to jointly optimize diacritic restoration with related NLP problems namely word segmentation, part-of-speech tagging, and syntactic diacritization. We use Arabic as a case study since it has sufficient data resources for tasks that we consider in our joint modeling. Our joint models significantly outperform the baselines and are comparable to the state-of-the-art models that are more complex relying on morphological analyzers and/or a lot more data (e.g. dialectal data).

Frugal Paradigm Completion [Website][PDF]
Alexander Erdmann, Tom Kenter, Markus Becker, and Christian Schallhart 17:00–18:00

Lexica distinguishing all morphologically related forms of each lexeme are crucial to many language technologies, yet building them is expensive. We propose a frugal paradigm completion approach that predicts all related forms in a morphological paradigm from as few manually provided forms as possible. It induces typological information during training which it uses to determine the best sources at test time. We evaluate our language-agnostic approach on 7 diverse languages. Compared to popular alternative approaches, ours reduces manual labor by 16-63% and is the most robust to typological variation.

Improving Chinese Word Segmentation with Wordhood Memory Networks [Website][PDF]
Yuanhe Tian, Yan Song, Fei Xia, Tong Zhang, and Yonggang Wang 17:00–18:00

Contextual features always play an important role in Chinese word segmentation (CWS). Wordhood information, being one of the contextual features, is proved to be useful in many conventional character-based segmenters. However, this feature receives less attention in recent neural models and it is also challenging to design a framework that can properly integrate wordhood information from different wordhood measures to existing neural frameworks. In this paper, we therefore propose a neural framework, WMSeg, which uses memory networks to incorporate wordhood information with several popular encoder-decoder combinations for CWS. Experimental results on five benchmark datasets indicate the memory mechanism successfully models wordhood information for neural segmenters and helps WMSeg achieve state-of-the-art performance on all those datasets. Further experiments and analyses also demonstrate the robustness of our proposed framework with respect to different wordhood measures and the efficiency of wordhood information in cross-domain experiments.

Joint Chinese Word Segmentation and Part-of-speech Tagging via Two-way Attentions of Auto-analyzed Knowledge [Website][PDF]
Yuanhe Tian, Yan Song, Xiang Ao, Fei Xia, Xiaojun Quan, Tong Zhang, and Yonggang Wang 17:00–18:00

Chinese word segmentation (CWS) and part-of-speech (POS) tagging are important fundamental tasks for Chinese language processing, where joint learning of them is an effective one-step solution for both tasks. Previous studies for joint CWS and POS tagging mainly follow the character-based tagging paradigm with introducing contextual information such as n-gram features or sentential representations from recurrent neural models. However, for many cases, the joint tagging needs not only modeling from context features but also knowledge attached to them (e.g., syntactic relations among words); limited efforts have been made by existing research to meet such needs. In this paper, we propose a neural model named TwASP for joint CWS and POS tagging following the character-based sequence labeling paradigm, where a two-way attention mechanism is used to incorporate both context feature and their corresponding syntactic knowledge for each input character. Particularly, we use existing language processing toolkits to obtain the auto-analyzed syntactic knowledge for the context, and the proposed attention module can learn and benefit from them although their quality may not be perfect. Our experiments illustrate the effectiveness of the two-way attentions for joint CWS and POS tagging, where state-of-the-art performance is achieved on five benchmark datasets.

Joint Diacritization, Lemmatization, Normalization, and Fine-Grained Morphological Tagging [Website][PDF]
Nasser Zalmout and Nizar Habash 17:00–18:00

The written forms of Semitic languages are both highly ambiguous and morphologically rich: a word can have multiple interpretations and is one of many inflected forms of the same concept or lemma. This is further exacerbated for dialectal content, which is more prone to noise and lacks a standard orthography. The morphological features can be lexicalized, like lemmas and diacritized forms, or non-lexicalized, like gender, number, and part-of-speech tags, among others. Joint modeling of the lexicalized and non-lexicalized features can identify more intricate morphological patterns, which provide better context modeling, and further disambiguate ambiguous lexical choices. However, the different modeling granularity can make joint modeling more difficult. Our approach models the different features jointly, whether lexicalized (on the character-level), or non-lexicalized (on the word-level). We use Arabic as a test case, and achieve state-of-the-art results for Modern Standard Arabic with 20% relative error reduction, and Egyptian Arabic with 11% relative error reduction.

Phonetic and Visual Priors for Decipherment of Informal Romanization [Website][PDF]
Maria Ryskina, Matthew R. Gormley, and Taylor Berg-Kirkpatrick 17:00–18:00

Informal romanization is an idiosyncratic process used by humans in informal digital communication to encode non-Latin script languages into Latin character sets found on common keyboards. Character substitution choices differ between users but have been shown to be governed by the same main principles observed across a variety of languages—namely, character pairs are often associated through phonetic or visual similarity. We propose a noisy-channel WFST cascade model for deciphering the original non-Latin script from observed romanized text in an unsupervised fashion. We train our model directly on romanized data from two languages: Egyptian Arabic and Russian. We demonstrate that adding inductive bias through phonetic and visual priors on character mappings substantially improves the model's performance on both languages, yielding results much closer to the supervised skyline. Finally, we introduce a new dataset of romanized Russian, collected from a Russian social network website and partially annotated for our experiments.

[TACL] Phonotactic Complexity and its Trade-offs [Website][PDF]
Tiago Pimentel, Brian Roark, and Ryan D. Cotterell 17:00–18:00

We present methods for calculating a measure of phonotactic complexity—bits per phoneme—that permits a straightforward cross-linguistic comparison. When given a word, represented as a sequence of phonemic segments such as symbols in the international phonetic alphabet, and a statistical model trained on a sample of word types from the language, we can approximately measure bits per phoneme using the negative log-probability of that word under the model. This simple measure allows us to compare the entropy across languages, giving insight into how complex a language's phonotactics are. Using a collection of 1016 basic concept words across 106 languages, we demonstrate a very strong negative correlation of -0.74 between bits per phoneme and the average length of words.

The Paradigm Discovery Problem [Website][PDF]
Alexander Erdmann, Micha Elsner, Shijie Wu, Ryan Cotterell, and Nizar Habash 17:00–18:00

This work treats the paradigm discovery problem (PDP), the task of learning an inflectional morphological system from unannotated sentences. We formalize the PDP and develop evaluation metrics for judging systems. Using currently available resources, we construct datasets for the task. We also devise a heuristic benchmark for the PDP and report empirical results on five diverse languages. Our benchmark system first makes use of word embeddings and string similarity to cluster forms by cell and by paradigm. Then, we bootstrap a neural transducer on top of the clustered data to predict words to realize the empty paradigm slots. An error analysis of our system suggests clustering by cell across different inflection classes is the most pressing challenge for future work.

Supervised Grapheme-to-Phoneme Conversion of Orthographic Schwas in Hindi and Punjabi [Website][PDF]
Aryaman Arora, Luke Gessler, and Nathan Schneider 17:00–18:00

Hindi grapheme-to-phoneme (G2P) conversion is mostly trivial, with one exception: whether a schwa represented in the orthography is pronounced or unpronounced (deleted). Previous work has attempted to predict schwa deletion in a rule-based fashion using prosodic or phonetic analysis. We present the first statistical schwa deletion classifier for Hindi, which relies solely on the orthography as the input and outperforms previous approaches. We trained our model on a newly-compiled pronunciation lexicon extracted from various online dictionaries. Our best Hindi model achieves state of the art performance, and also achieves good performance on a closely related language, Punjabi, without modification.

Session 14A Semantics: Sentence Level-9

Active Learning for Coreference Resolution using Discrete Annotation [Website][PDF]
Belinda Z. Li, Gabriel Stanovsky, and Luke Zettlemoyer 17:00–18:00

We improve upon pairwise annotation for active learning in coreference resolution, by asking annotators to identify mention antecedents if a presented mention pair is deemed not coreferent. This simple modification, when combined with a novel mention clustering algorithm for selecting which examples to label, is much more efficient in terms of the performance obtained per annotation budget. In experiments with existing benchmark coreference datasets, we show that the signal from this additional question leads to significant performance gains per human-annotation hour. Future work can use our annotation protocol to effectively develop coreference models for new domains. Our code is publicly available.

Beyond Possession Existence: Duration and Co-Possession [Website][PDF]
Dhivya Chinnappa, Srikala Murugan, and Eduardo Blanco 17:00–18:00

This paper introduces two tasks: determining (a) the duration of possession relations and (b) co-possessions, i.e., whether multiple possessors possess a possessee at the same time. We present new annotations on top of corpora annotating possession existence and experimental results. Regarding possession duration, we derive the time spans we work with empirically from annotations indicating lower and upper bounds. Regarding co-possessions, we use a binary label. Cohen's kappa coefficients indicate substantial agreement, and experimental results show that text is more useful than the image for solving these tasks.

[TACL] Decoding Brain Activity Associated with Literal and Metaphoric Sentence Comprehension using Distributional Semantic Models [Website][PDF]
Vesna G. Djokic, Jean Maillard, Luana Bulat, and Ekaterina Shutova 17:00–18:00

Recent years have seen a growing interest within the natural language processing (NLP) community in evaluating the ability of semantic models to capture human meaning representation in the brain. Existing research has mainly focused on applying semantic models to decode brain activity patterns associated with the meaning of individual words, and, more recently, this approach has been extended to sentences and larger text fragments. Our work is the first to investigate metaphor processing in the brain in this context. We evaluate a range of semantic models (word embeddings, compositional, and visual models) in their ability to decode brain activity associated with reading of both literal and metaphoric sentences. Our results suggest that compositional models and word embeddings are able to capture differences in the processing of literal and metaphoric sentences, providing support for the idea that the literal meaning is not fully accessible during familiar metaphor comprehension.

Don't Stop Pretraining: Adapt Language Models to Domains and Tasks [Website][PDF]
Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith 17:00–18:00

Language models pretrained on text from a wide variety of sources form the foundation of today's NLP. In light of the success of these broad-coverage models, we investigate whether it is still helpful to tailor a pretrained model to the domain of a target task. We present a study across four domains (biomedical and computer science publications, news, and reviews) and eight classification tasks, showing that a second phase of pretraining in-domain (domain-adaptive pretraining) leads to performance gains, under both high- and low-resource settings. Moreover, adapting to the task's unlabeled data (task-adaptive pretraining) improves performance even after domain-adaptive pretraining. Finally, we show that adapting to a task corpus augmented using simple data selection strategies is an effective alternative, especially when resources for domain-adaptive pretraining might be unavailable. Overall, we consistently find that multi-phase adaptive pretraining offers large gains in task performance.

Estimating Mutual Information Between Dense Word Embeddings [Website][PDF]
Vitalii Zhelezniak, Aleksandar Savkov, and Nils Hammerla 17:00–18:00

Word embedding-based similarity measures are currently among the top-performing methods on unsupervised semantic textual similarity (STS) tasks. Recent work has increasingly adopted a statistical view on these embeddings, with some of the top approaches being essentially various correlations (which include the famous cosine similarity). Another excellent candidate for a similarity measure is mutual information (MI), which can capture arbitrary dependencies between the variables and has a simple and intuitive expression. Unfortunately, its use in the context of dense word embeddings has so far been avoided due to difficulties with estimating MI for continuous data. In this work we go through a vast literature on estimating MI in such cases and single out the most promising methods, yielding a simple and elegant similarity measure for word embeddings. We show that mutual information is a viable alternative to correlations, gives an excellent signal that correlates well with human judgements of similarity and rivals existing state-of-the-art unsupervised methods.

Exploring Unexplored Generalization Challenges for Cross-Database Semantic Parsing [Website][PDF]
Alane Suhr, Ming-Wei Chang, Peter Shaw, and Kenton Lee 17:00–18:00

We study the task of cross-database semantic parsing (XSP), where a system that maps natural language utterances to executable SQL queries is evaluated on databases unseen during training. Recently, several datasets, including Spider, were proposed to support development of XSP systems. We propose a challenging evaluation setup for cross-database semantic parsing, focusing on variation across database schemas and in-domain language use. We re-purpose eight semantic parsing datasets that have been well-studied in the setting where in-domain training data is available, and

instead use them as additional evaluation data for XSP systems instead. We build a system that performs well on Spider, and find that it struggles to generalize to our re-purposed set. Our setup uncovers several generalization challenges for cross-database semantic parsing, demonstrating the need to use and develop diverse training and evaluation datasets.

Predicting the Focus of Negation: Model and Error Analysis [Website][PDF]
Md Mosharaf Hossain, Kathleen Hamilton, Alexis Palmer, and Eduardo Blanco 17:00–18:00

The focus of a negation is the set of tokens intended to be negated, and a key component for revealing affirmative alternatives to negated utterances. In this paper, we experiment with neural networks to predict the focus of negation. Our main novelty is leveraging a scope detector to introduce the scope of negation as an additional input to the network. Experimental results show that doing so obtains the best results to date. Additionally, we perform a detailed error analysis providing insights into the main error categories, and analyze errors depending on whether the model takes into account scope and context information.

RAT-SQL: Relation-Aware Schema Encoding and Linking for Text-to-SQL Parsers [Website][PDF]
Bailin Wang, Richard Shin, Xiaodong Liu, Oleksandr Polozov, and Matthew Richardson 17:00–18:00

When translating natural language questions into SQL queries to answer questions from a database, contemporary semantic parsing models struggle to generalize to unseen database schemas. The generalization challenge lies in (a) encoding the database relations in an accessible way for the semantic parser, and (b) modeling alignment between database columns and their mentions in a given query. We present a unified framework, based on the relation-aware self-attention mechanism, to address schema encoding, schema linking, and feature representation within a text-to-SQL encoder. On the challenging Spider dataset this framework boosts the exact match accuracy to 57.2%, surpassing its best counterparts by 8.7% absolute improvement. Further augmented with BERT, it achieves the new state-of-the-art performance of 65.6% on the Spider leaderboard. In addition, we observe qualitative improvements in the model's understanding of schema linking and alignment. Our implementation will be open-sourced at https://github.com/Microsoft/rat-sql.

Structured Tuning for Semantic Role Labeling [Website][PDF]
Tao Li, Parth Anand Jawale, Martha Palmer, and Vivek Srikumar 17:00–18:00

Recent neural network-driven semantic role labeling (SRL) systems have shown impressive improvements in F1 scores. These improvements are due to expressive input representations, which, at least at the surface, are orthogonal to knowledge-rich constrained decoding mechanisms that helped linear SRL models. Introducing the benefits of structure to inform neural models presents a methodological challenge. In this paper, we present a structured tuning framework to improve models using softened constraints only at training time. Our framework leverages the expressiveness of neural networks and provides supervision with structured loss components. We start with a strong baseline (RoBERTa) to validate the impact of our approach, and show that our framework outperforms the baseline by learning to comply with declarative constraints. Additionally, our experiments with smaller training sizes show that we can achieve consistent improvements under low-resource scenarios.

TaBERT: Pretraining for Joint Understanding of Textual and Tabular Data [Website][PDF]
Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel 17:00–18:00

Recent years have witnessed the burgeoning of pretrained language models (LMs) for text-based natural language (NL) understanding tasks. Such models are typically trained on free-form NL text, hence may not be suitable for tasks like semantic parsing over structured data, which require reasoning over both free-form NL questions and structured tabular data (e.g., database tables). In this paper we present TaBERT, a pretrained LM that jointly learns representations for NL sentences and (semi-)structured tables. TaBERT is trained on a large corpus of 26 million tables and their English contexts. In experiments, neural semantic parsers using TaBERT as feature representation layers achieve new best results on the challenging weakly-supervised semantic parsing benchmark WikiTableQuestions, while performing competitively on the text-to-SQL dataset Spider.

Universal Decompositional Semantic Parsing [Website][PDF]
Elias Stengel-Eskin, Aaron Steven White, Sheng Zhang, and Benjamin Van Durme 17:00–18:00

We introduce a transductive model for parsing into Universal Decompositional Semantics (UDS) representations, which jointly learns to map natural language utterances into UDS graph structures and annotate the graph with decompositional semantic attribute scores. We also introduce a strong pipeline model for parsing into the UDS graph structure, and show that our transductive parser performs comparably while additionally performing attribute prediction. By analyzing the attribute prediction errors, we find the model captures natural relationships between attribute groups.

Unsupervised Cross-lingual Representation Learning at Scale [Website][PDF]
Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov 17:00–18:00

This paper shows that pretraining multilingual language models at scale leads to significant performance gains for a wide range of cross-lingual transfer tasks. We train a Transformer-based masked language model on one hundred languages, using more than two terabytes of filtered CommonCrawl data. Our model, dubbed XLM-R, significantly outperforms multilingual BERT (mBERT) on a variety of cross-lingual benchmarks, including +14.6% average accuracy on XNLI, +13% average F1 score on MLQA, and +2.4% F1 score on NER. XLM-R performs particularly well on low-resource languages, improving 15.7% in XNLI accuracy for Swahili and 11.4% for Urdu over previous XLM models. We also present a detailed empirical analysis of the key factors that are required to achieve these gains, including the trade-offs between (1) positive transfer and capacity dilution and (2) the performance of high and low resource

languages at scale. Finally, we show, for the first time, the possibility of multilingual modeling without sacrificing per-language performance; XLM-R is very competitive with strong monolingual models on the GLUE and XNLI benchmarks. We will make our code and models publicly available.

Session 14A: Student Research Workshop

Topic Balancing with Additive Regularization of Topic Models [Website][PDF]
Eugeniia Veselova and Konstantin Vorontsov 17:00–18:00

This article proposes a new approach for building topic models on unbalanced collections in topic modelling, based on the existing methods and our experiments with such methods. Real-world data collections contain topics in various proportions, and often documents of the relatively small theme become distributed all over the larger topics instead of being grouped into one topic. To address this issue, we design a new regularizer for Theta and Phi matrices in probabilistic Latent Semantic Analysis (pLSA) model. We make sure this regularizer increases the quality of topic models, trained on unbalanced collections. Besides, we conceptually support this regularizer by our experiments.

Combining Subword Representations into Word-level Representations in the Transformer Architecture [Website][PDF]
Noe Casas, Marta R. Costa-jussà, and José A. R. Fonollosa 17:00–18:00

In Neural Machine Translation, using word-level tokens leads to degradation in translation quality. The dominant approaches use subword-level tokens, but this increases the length of the sequences and makes it difficult to profit from word-level information such as POS tags or semantic dependencies. We propose a modification to the Transformer model to combine subword-level representations into word-level ones in the first layers of the encoder, reducing the effective length of the sequences in the following layers and providing a natural point to incorporate extra word-level information. Our experiments show that this approach maintains the translation quality with respect to the normal Transformer model when no extra word-level information is injected and that it is superior to the currently dominant method for incorporating word-level source language information to models based on subword-level vocabularies.

Exploring Interpretability in Event Extraction: Multitask Learning of a Neural Event Classifier and an Explanation Decoder [Website][PDF]
Zheng Tang, Gus Hahn-Powell, and Mihai Surdeanu 17:00–18:00

We propose an interpretable approach for event extraction that mitigates the tension between generalization and interpretability by jointly training for the two goals. Our approach uses an encoder-decoder architecture, which jointly trains a classifier for event extraction, and a rule decoder that generates syntactico-semantic rules that explain the decisions of the event classifier. We evaluate the proposed approach on three biomedical events and show that the decoder generates interpretable rules that serve as accurate explanations for the event classifier's decisions, and, importantly, that the joint training generally improves the performance of the event classifier. Lastly, we show that our approach can be used for semi-supervised learning, and that its performance improves when trained on automatically-labeled data generated by a rule-based system.

Dominance as an Indicator of Rapport and Learning in Human-Agent Communication [Website]
Amanda Buddemeyer, Xiaoyi Tian, and Erin Walker 17:00–18:00

Power dynamics in human-human communication can impact rapport-building and learning gains, but little is known about how power impacts human-agent communication. In this paper, we examine dominance behavior in utterances between middle-school students and a teachable robot as they work through math problems, as coded by Rogers and Farace's Relational Communication Control Coding Scheme (RCCCS). We hypothesize that relatively dominant students will show increased learning gains, as will students with greater dominance agreement with the robot. We also hypothesize that gender could be an indicator of differences in dominance behavior. We present a preliminary analysis of dominance characteristics in some of the transactions between robot and student. Ultimately, we hope to determine if manipulating the dominance behavior of a learning robot could support learning.

Demo Session 4B

Time: *17:45–18:30*

Conversation Learner - A Machine Teaching Tool for Building Dialog Managers for Task-Oriented Dialog Systems [Website][PDF]

Swadheen Shukla, Lars Liden, Shahin Shayandeh, Eslam Kamal, Jinchao Li, Matt Mazzola, Thomas Park, Baolin Peng, and Jianfeng Gao

Traditionally, industry solutions for building a task-oriented dialog system have relied on helping dialog authors define rule-based dialog managers, represented as dialog flows. While dialog flows are intuitively interpretable and good for simple scenarios, they fall short of performance in terms of the flexibility needed to handle complex dialogs. On the other hand, purely machine-learned models can handle complex dialogs, but they are considered to be black boxes and require large amounts of training data. In this demonstration, we showcase Conversation Learner, a machine teaching tool for building dialog managers. It combines the best of both approaches by enabling dialog authors to create a dialog flow using familiar tools, converting the dialog flow into a parametric model (e.g., neural networks), and allowing dialog authors to improve the dialog manager (i.e., the parametric model) over time by leveraging user-system dialog logs as training data through a machine teaching interface.

LEAN-LIFE: A Label-Efficient Annotation Framework Towards Learning from Explanation [Website][PDF]

Dong-Ho Lee, Rahul Khanna, Bill Yuchen Lin, Seyeon Lee, Qinyuan Ye, Elizabeth Boschee, Leonardo Neves, and Xiang Ren

Successfully training a deep neural network demands a huge corpus of labeled data. However, each label only provides limited information to learn from, and collecting the requisite number of labels involves massive human effort. In this work, we introduce LEAN-LIFE, a web-based, Label-Efficient AnnotatioN framework for sequence labeling and classification tasks, with an easy-to-use UI that not only allows an annotator to provide the needed labels for a task but also enables LearnIng From Explanations for each labeling decision. Such explanations enable us to generate useful additional labeled data from unlabeled instances, bolstering the pool of available training data. On three popular NLP tasks (named entity recognition, relation extraction, sentiment analysis), we find that using this enhanced supervision allows our models to surpass competitive baseline F1 scores by more than 5-10 percentage points, while using 2X times fewer labeled instances. Our framework is the first to utilize this enhanced supervision technique and does so for three important tasks – thus providing improved annotation recommendations to users and an ability to build datasets of (data, label, explanation) triples instead of the regular (data, label) pair.

Session 14B Overview – Wednesday, July 8, 2020 18:00–19:00

Track A *Information Extraction-10* Abstracts	A Generate-and-Rank Framework with Semantic Type Regularization for Biomedical Concept Normalization *Xu, Zhang, and Bethard* [Website][PDF]	Hierarchical Entity Typing via Multi-level Learning to Rank *Chen, Chen, and Van Durme* [Website][PDF]	Multi-Domain Named Entity Recognition with Genre-Aware and Agnostic Inference *Wang, Kulkarni, and Preotiuc-Pietro* [Website][PDF]	TXtract: Taxonomy-Aware Knowledge Extraction for Thousands of Product Categories *Karamanolakis, Ma, and Dong* [Website][PDF]	TriggerNER: Learning with Entity Triggers as Explanations for Named Entity Recognition *Lin, Lee, Shen, Moreno, Huang, Shiralkar, and Ren* [Website][PDF]
Track B *Machine Learning for NLP-16* Abstracts	A Mixture of h - 1 Heads is Better than h Heads *Peng, Schwartz, Li, and Smith* [Website][PDF]	CamemBERT: a Tasty French Language Model *Martin, Muller, Ortiz Suárez, Dupont, Romary, Clergerie, Seddah, and Sagot* [Website][PDF]	Contrastive Self-Supervised Learning for Commonsense Reasoning *Klein and Nabi* [Website][PDF]	Do Transformers Need Deep Long-Range Memory? *Rae and Razavi* [Website][PDF]	Generalized Entropy Regularization or: There's Nothing Special about Label Smoothing *Meister, Salesky, and Cotterell* [Website][PDF]
	Improving Disentangled Text Representation Learning with Information-Theoretic Guidance *Cheng, Min, Shen, Malon, Zhang, Li, and Carin* [Website][PDF]	The Right Tool for the Job: Matching Model and Instance Complexities *Schwartz, Stanovsky, Swayamdipta, Dodge, and Smith* [Website][PDF]			
Track C *Machine Translation-16* Abstracts	Addressing Posterior Collapse with Mutual Information for Improved Variational Neural Machine Translation *McCarthy, Li, Gu, and Dong* [Website][PDF]	Balancing Training for Multilingual Neural Machine Translation *Wang, Tsvetkov, and Neubig* [Website][PDF]	Classification-Based Self-Learning for Weakly Supervised Bilingual Lexicon Induction *Karan, Vulić, Korhonen, and Glavaš* [Website][PDF]	Evaluating Robustness to Input Perturbations for Neural Machine Translation *Niu, Mathur, Dinu, and Al-Onaizan* [Website][PDF]	Gender in Danger? Evaluating Speech Translation Technology on the MuST-SHE Corpus *Bentivogli, Savoldi, Negri, Di Gangi, Cattoni, and Turchi* [Website][PDF]
	In Neural Machine Translation, What Does Transfer Learning Transfer? *Aji, Bogoychev, Heafield, and Sennrich* [Website][PDF]	Parallel Corpus Filtering via Pre-trained Language Models *Zhang, Nagesh, and Knight* [Website][PDF]	Reducing Gender Bias in Neural Machine Translation as a Domain Adaptation Problem *Saunders and Byrne* [Website][PDF]	Regularized Context Gates on Transformer for Machine Translation *Li, Liu, Wang, Huang, and Meng* [Website][PDF]	Unsupervised Domain Clusters in Pretrained Language Models *Aharoni and Goldberg* [Website][PDF]
	Using Context in Neural Machine Translation Training Objectives *Saunders, Stahlberg, and Byrne* [Website][PDF]	Variational Neural Machine Translation with Normalizing Flows *Setiawan, Sperber, Nallasamy, and Paulik* [Website][PDF]			

Track D *NLP Applications-11* Abstracts	A Multi-Perspective Architecture for Semantic Code Search *Haldar, Wu, Xiong, and Hockenmaier* [Website][PDF]	Automated Topical Component Extraction Using Neural Network Attention Scores from Source-based Essay Scoring *Zhang and Litman* [Website][PDF]	Clinical Concept Linking with Contextualized Neural Representations *Schumacher, Mulyar, and Dredze* [Website][PDF]	DeSePtion: Dual Sequence Prediction and Adversarial Examples for Improved Fact-Checking *Hidey, Chakrabarty, Alhindi, Varia, Krstovski, Diab, and Muresan* [Website][PDF]	Estimating predictive uncertainty for rumour verification models *Kochkina and Liakata* [Website][PDF]
	Let Me Choose: From Verbal Context to Font Selection *Shirani, Dernoncourt, Echevarria, Asente, Lipka, and Solorio* [Website][PDF]	[TACL] Machine Learning Driven Language Assessment *Settles, Hagiwara, and LaFlair* [Website][PDF]	Multi-Label and Multilingual News Framing Analysis *Akyürek, Guo, Elanwar, Ishwar, Betke, and Wijaya* [Website][PDF]	Predicting Performance for Natural Language Processing Tasks *Xia, Anastasopoulos, Xu, Yang, and Neubig* [Website][PDF]	ScriptWriter: Narrative-Guided Script Generation *Zhu, Song, Dou, NIE, and Zhou* [Website][PDF]
	Should All Cross-Lingual Embeddings Speak English? *Anastasopoulos and Neubig* [Website][PDF]	Smart To-Do: Automatic Generation of To-Do Items from Emails *Mukherjee, Mukherjee, Hasegawa, Hassan Awadallah, and White* [Website][PDF]	Understanding Advertisements with BERT *Kalra, Kurma, Vadakkeveetil Sreelatha, Patwardhan, and Karande* [Website][PDF]		
Track E *Lexical-8* Abstracts	[CL] LESSLEX: Linking Multilingual Embeddings to SenSe Representations of Lexical Items *Colla, Mensa, and Radicioni* [Website][PDF]	Non-Linear Instance-Based Cross-Lingual Mapping for Non-Isomorphic Embedding Spaces *Glavaš and Vulić* [Website][PDF]			
Track F *Textual Inference and Other Areas of Semantics-6* Abstracts	Are Natural Language Inference Models IMPPRESsive? Learning IMPlicature and PRESupposition *Jeretic, Warstadt, Bhooshan, and Williams* [Website][PDF]	Benchmarking Multimodal Regex Synthesis with Complex Structures *Ye, Chen, Dillig, and Durrett* [Website][PDF]	End-to-End Bias Mitigation by Modelling Biases in Corpora *Karimi Mahabadi, Belinkov, and Henderson* [Website][PDF]	Generating Fact Checking Explanations *Atanasova, Simonsen, Lioma, and Augenstein* [Website][PDF]	Mind the Trade-off: Debiasing NLU Models without Degrading the In-distribution Performance *Utama, Moosavi, and Gurevych* [Website][PDF]
	NILE : Natural Language Inference with Faithful Natural Language Explanations *Kumar and Talukdar* [Website][PDF]	Neural Mixed Counting Models for Dispersed Topic Discovery *Wu, Rao, Zhang, Xie, Li, Wang, and Chen* [Website][PDF]	QuASE: Question-Answer Driven Sentence Encoding *He, Ning, and Roth* [Website][PDF]	Temporal Common Sense Acquisition with Minimal Supervision *Zhou, Ning, Khashabi, and Roth* [Website][PDF]	The Sensitivity of Language Models and Humans to Winograd Schema Perturbations *Abdou, Ravishankar, Barrett, Belinkov, Elliott, and Søgaard* [Website][PDF]
	Towards Robustifying NLI Models Against Lexical Dataset Biases *Zhou and Bansal* [Website][PDF]	Uncertain Natural Language Inference *Chen, Jiang, Poliak, Sakaguchi, and Van Durme* [Website][PDF]			

Track G *Student Research Workshop* Abstracts	Why is penguin more similar to polar bear than to sea gull? Analyzing conceptual knowledge in distributional models *Sommerauer* [Website][PDF]	Pointwise Paraphrase Appraisal is Potentially Problematic *Chen, Ji, and Evans* [Website][PDF]	A Geometry-Inspired Attack for Generating Natural Language Adversarial Examples *Meng and Wattenhofer* [Website]	Enhancing Word Embeddings with Knowledge Extracted from Lexical Resources *Biesialska, Rafieian, and Costa-jussà* [Website][PDF]	
Track H *Tagging, Chunking and Parsing-5* Abstracts	[TACL] Deep Contextualized Self-training for Low Resource Dependency Parsing *Rotman and Reichart* [Website][PDF]	Extracting Headless MWEs from Dependency Parse Trees: Parsing, Tagging, and Joint Modeling Approaches *Shi and Lee* [Website][PDF]	Revisiting Higher-Order Dependency Parsers *Fonseca and Martins* [Website][PDF]	SeqVAT: Virtual Adversarial Training for Semi-Supervised Sequence Labeling *Chen, Ruan, Liu, and Lu* [Website][PDF]	Tetra-Tagging: Word-Synchronous Parsing with Linear-Time Inference *Kitaev and Klein* [Website][PDF]
	Treebank Embedding Vectors for Out-of-domain Dependency Parsing *Wagner, Barry, and Foster* [Website][PDF]				

Session 14B Details

Session 14B: Information Extraction-10

A Generate-and-Rank Framework with Semantic Type Regularization for Biomedical Concept Normalization [Website][PDF]
Dongfang Xu, Zeyu Zhang, and Steven Bethard 18:00–19:00

Concept normalization, the task of linking textual mentions of concepts to concepts in an ontology, is challenging because ontologies are large. In most cases, annotated datasets cover only a small sample of the concepts, yet concept normalizers are expected to predict all concepts in the ontology. In this paper, we propose an architecture consisting of a candidate generator and a list-wise ranker based on BERT. The ranker considers pairings of concept mentions and candidate concepts, allowing it to make predictions for any concept, not just those seen during training. We further enhance this list-wise approach with a semantic type regularizer that allows the model to incorporate semantic type information from the ontology during training. Our proposed concept normalization framework achieves state-of-the-art performance on multiple datasets.

Hierarchical Entity Typing via Multi-level Learning to Rank [Website][PDF]
Tongfei Chen, Yunmo Chen, and Benjamin Van Durme 18:00–19:00

We propose a novel method for hierarchical entity classification that embraces ontological structure at both training and during prediction. At training, our novel multi-level learning-to-rank loss compares positive types against negative siblings according to the type tree. During prediction, we define a coarse-to-fine decoder that restricts viable candidates at each level of the ontology based on already predicted parent type(s). Our approach significantly outperform prior work on strict accuracy, demonstrating the effectiveness of our method.

Multi-Domain Named Entity Recognition with Genre-Aware and Agnostic Inference [Website][PDF]
Jing Wang, Mayank Kulkarni, and Daniel Preotiuc-Pietro 18:00–19:00

Named entity recognition is a key component of many text processing pipelines and it is thus essential for this component to be robust to different types of input. However, domain transfer of NER models with data from multiple genres has not been widely studied. To this end, we conduct NER experiments in three predictive setups on data from: a) multiple domains; b) multiple domains where the genre label is unknown at inference time; c) domains not encountered in training. We introduce a new architecture tailored to this task by using shared and private domain parameters and multi-task learning. This consistently outperforms all other baseline and competitive methods on all three experimental setups, with differences ranging between +1.95 to +3.11 average F1 across multiple genres when compared to standard approaches. These results illustrate the challenges that need to be taken into account when building real-world NLP applications that are robust to various types of text and the methods that can help, at least partially, alleviate these issues.

TXtract: Taxonomy-Aware Knowledge Extraction for Thousands of Product Categories [Website][PDF]
Giannis Karamanolakis, Jun Ma, and Xin Luna Dong 18:00–19:00

Extracting structured knowledge from product profiles is crucial for various applications in e-Commerce. State-of-the-art approaches for knowledge extraction were each designed for a single category of product, and thus do not apply to real-life e-Commerce scenarios, which often contain thousands of diverse categories. This paper proposes TXtract, a taxonomy-aware knowledge extraction model that applies to thousands of product categories organized in a hierarchical taxonomy. Through category conditional self-attention and multi-task learning, our approach is both scalable, as it trains a single model for thousands of categories, and effective, as it extracts category-specific attribute values. Experiments on products from a taxonomy with 4,000 categories show that TXtract outperforms state-of-the-art approaches by up to 10% in F1 and 15% in coverage across all categories.

TriggerNER: Learning with Entity Triggers as Explanations for Named Entity Recognition [Website][PDF]
Bill Yuchen Lin, Dong-Ho Lee, Ming Shen, Ryan Moreno, Xiao Huang, Prashant Shiralkar, and Xiang Ren 18:00–19:00

Training neural models for named entity recognition (NER) in a new domain often requires additional human annotations (e.g., tens of thousands of labeled instances) that are usually expensive and time-consuming to collect. Thus, a crucial research question is how to obtain supervision in a cost-effective way. In this paper, we introduce "entity triggers," an effective proxy of human explanations for facilitating label-efficient learning of NER models. An entity trigger is defined as a group of words in a sentence that helps to explain why humans would recognize an entity in the sentence. We crowd-sourced 14k entity triggers for two well-studied NER datasets. Our proposed model, Trigger Matching Network, jointly learns trigger representations and soft matching module with self-attention such that can generalize to unseen sentences easily for tagging. Our framework is significantly more cost-effective than the traditional neural NER frameworks. Experiments show that using only 20% of the trigger-annotated sentences results in a comparable performance as using 70% of conventional annotated sentences.

Session 14B: Machine Learning for NLP-16

A Mixture of h - 1 Heads is Better than h Heads [Website][PDF]
Hao Peng, Roy Schwartz, Dianqi Li, and Noah A. Smith 18:00–19:00

Multi-head attentive neural architectures have achieved state-of-the-art results on a variety of natural language processing tasks. Evidence has shown that they are overparameterized; attention heads can be pruned without significant performance loss. In this work, we instead "reallocate" them—the model learns to activate different heads on different inputs. Drawing connections between multi-head attention and mixture of experts, we propose the mixture of attentive experts model (MAE). MAE is trained using a block coordinate descent algorithm that alternates between updating (1) the responsibilities of the experts and (2) their parameters. Experiments on machine translation and language modeling show that MAE outperforms strong baselines on both tasks. Particularly, on the WMT14 English to German translation dataset, MAE improves over "transformer-base" by 0.8 BLEU, with a comparable number of parameters. Our analysis shows that our model learns to specialize different experts to different inputs.

CamemBERT: a Tasty French Language Model [Website][PDF]
Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot 18:00–19:00

Pretrained language models are now ubiquitous in Natural Language Processing. Despite their success, most available models have either been trained on English data or on the concatenation of data in multiple languages. This makes practical use of such models –in all languages except English– very limited. In this paper, we investigate the feasibility of training monolingual Transformer-based language models for other languages, taking French as an example and evaluating our language models on part-of-speech tagging, dependency parsing, named entity recognition and natural language inference tasks. We show that the use of web crawled data is preferable to the use of Wikipedia data. More surprisingly, we show that a relatively small web crawled dataset (4GB) leads to results that are as good as those obtained using larger datasets (130+GB). Our best performing model CamemBERT reaches or improves the state of the art in all four downstream tasks.

Contrastive Self-Supervised Learning for Commonsense Reasoning [Website][PDF]
Tassilo Klein and Moin Nabi 18:00–19:00

We propose a self-supervised method to solve Pronoun Disambiguation and Winograd Schema Challenge problems. Our approach exploits the characteristic structure of training corpora related to so-called "trigger" words, which are responsible for flipping the answer in pronoun disambiguation. We achieve such commonsense reasoning by constructing pair-wise contrastive auxiliary predictions. To this end, we leverage a mutual exclusive loss regularized by a contrastive margin. Our architecture is based on the recently introduced transformer networks, BERT, that exhibits strong performance on many NLP benchmarks. Empirical results show that our method alleviates the limitation of current supervised approaches for commonsense reasoning. This study opens up avenues for exploiting inexpensive self-supervision to achieve performance gain in commonsense reasoning tasks.

Do Transformers Need Deep Long-Range Memory? [Website][PDF]
Jack Rae and Ali Razavi 18:00–19:00

Deep attention models have advanced the modelling of sequential data across many domains. For language modelling in particular, the Transformer-XL — a Transformer augmented with a long-range memory of past activations — has been shown to be state-of-the-art across a variety of well-studied benchmarks. The Transformer-XL incorporates a long-range memory at every layer of the network, which renders its state to be thousands of times larger than RNN predecessors. However it is unclear whether this is necessary. We perform a set of interventions to show that comparable performance can be obtained with 6X fewer long range memories and better performance can be obtained by limiting the range of attention in lower layers of the network.

Generalized Entropy Regularization or: There's Nothing Special about Label Smoothing [Website][PDF]
Clara Meister, Elizabeth Salesky, and Ryan Cotterell 18:00–19:00

Prior work has explored directly regularizing the output distributions of probabilistic models to alleviate peaky (i.e. over-confident) predictions, a common sign of overfitting. This class of techniques, of which label smoothing is one, has a connection to entropy regularization. Despite the consistent success of label smoothing across architectures and data sets in language generation tasks, two problems remain open: (1) there is little understanding of the underlying effects entropy regularizers have on models, and (2) the full space of entropy regularization techniques is largely unexplored. We introduce a parametric family of entropy regularizers, which includes label smoothing as a special case, and use it to gain a better understanding of the relationship between the entropy of a model and its performance on language generation tasks. We also find that variance in model performance can be explained largely by the resulting entropy of the model. Lastly, we find that label smoothing provably does not allow for sparsity in an output distribution, an undesirable property for language generation models, and therefore advise the use of other entropy regularization methods in its place.

Improving Disentangled Text Representation Learning with Information-Theoretic Guidance [Website][PDF]
Pengyu Cheng, Martin Renqiang Min, Dinghan Shen, Christopher Malon, Yizhe Zhang, Yitong Li, and Lawrence Carin 18:00–19:00

Learning disentangled representations of natural language is essential for many NLP tasks, e.g., conditional text gen-

eration, style transfer, personalized dialogue systems, etc. Similar problems have been studied extensively for other forms of data, such as images and videos. However, the discrete nature of natural language makes the disentangling of textual representations more challenging (e.g., the manipulation over the data space cannot be easily achieved). Inspired by information theory, we propose a novel method that effectively manifests disentangled representations of text, without any supervision on semantics. A new mutual information upper bound is derived and leveraged to measure dependence between style and content. By minimizing this upper bound, the proposed method induces style and content embeddings into two independent low-dimensional spaces. Experiments on both conditional text generation and text-style transfer demonstrate the high quality of our disentangled representation in terms of content and style preservation.

The Right Tool for the Job: Matching Model and Instance Complexities [Website] [PDF]
Roy Schwartz, Gabriel Stanovsky, Swabha Swayamdipta, Jesse Dodge, and Noah A. Smith 18:00–19:00

As NLP models become larger, executing a trained model requires significant computational resources incurring monetary and environmental costs. To better respect a given inference budget, we propose a modification to contextual representation fine-tuning which, during inference, allows for an early (and fast) "exit" from neural network calculations for simple instances, and late (and accurate) exit for hard instances. To achieve this, we add classifiers to different layers of BERT and use their calibrated confidence scores to make early exit decisions. We test our proposed modification on five different datasets in two tasks: three text classification datasets and two natural language inference benchmarks. Our method presents a favorable speed/accuracy tradeoff in almost all cases, producing models which are up to five times faster than the state of the art, while preserving their accuracy. Our method also requires almost no additional training resources (in either time or parameters) compared to the baseline BERT model. Finally, our method alleviates the need for costly retraining of multiple models at different levels of efficiency; we allow users to control the inference speed/accuracy tradeoff using a single trained model, by setting a single variable at inference time. We publicly release our code.

Session 14B: Machine Translation-16

Addressing Posterior Collapse with Mutual Information for Improved Variational Neural Machine Translation [Website][PDF]
Arya D. McCarthy, Xian Li, Jiatao Gu, and Ning Dong 18:00–19:00

This paper proposes a simple and effective approach to address the problem of posterior collapse in conditional variational autoencoders (CVAEs). It thus improves performance of machine translation models that use noisy or monolingual data, as well as in conventional settings. Extending Transformer and conditional VAEs, our proposed latent variable model measurably prevents posterior collapse by (1) using a modified evidence lower bound (ELBO) objective which promotes mutual information between the latent variable and the target, and (2) guiding the latent variable with an auxiliary bag-of-words prediction task. As a result, the proposed model yields improved translation quality compared to existing variational NMT models on WMT Ro↔En and De↔En. With latent variables being effectively utilized, our model demonstrates improved robustness over non-latent Transformer in handling uncertainty: exploiting noisy source-side monolingual data (up to +3.2 BLEU), and training with weakly aligned web-mined parallel data (up to +4.7 BLEU).

Balancing Training for Multilingual Neural Machine Translation [Website][PDF]
Xinyi Wang, Yulia Tsvetkov, and Graham Neubig 18:00–19:00

When training multilingual machine translation (MT) models that can translate to/from multiple languages, we are faced with imbalanced training sets: some languages have much more training data than others. Standard practice is to up-sample less resourced languages to increase representation, and the degree of up-sampling has a large effect on the overall performance. In this paper, we propose a method that instead automatically learns how to weight training data through a data scorer that is optimized to maximize performance on all test languages. Experiments on two sets of languages under both one-to-many and many-to-one MT settings show our method not only consistently outperforms heuristic baselines in terms of average performance, but also offers flexible control over the performance of which languages are optimized.

Classification-Based Self-Learning for Weakly Supervised Bilingual Lexicon Induction [Website][PDF]
Mladen Karan, Ivan Vulić, Anna Korhonen, and Goran Glavaš 18:00–19:00

Effective projection-based cross-lingual word embedding (CLWE) induction critically relies on the iterative self-learning procedure. It gradually expands the initial small seed dictionary to learn improved cross-lingual mappings. In this work, we present ClassyMap, a classification-based approach to self-learning, yielding a more robust and a more effective induction of projection-based CLWEs. Unlike prior self-learning methods, our approach allows for integration of diverse features into the iterative process. We show the benefits of ClassyMap for bilingual lexicon induction: we report consistent improvements in a weakly supervised setup (500 seed translation pairs) on a benchmark with 28 language pairs.

Evaluating Robustness to Input Perturbations for Neural Machine Translation [Website][PDF]
Xing Niu, Prashant Mathur, Georgiana Dinu, and Yaser Al-Onaizan 18:00–19:00

Neural Machine Translation (NMT) models are sensitive to small perturbations in the input. Robustness to such perturbations is typically measured using translation quality metrics such as BLEU on the noisy input. This paper proposes additional metrics which measure the relative degradation and changes in translation when small perturbations are added to the input. We focus on a class of models employing subword regularization to address robustness and perform extensive evaluations of these models using the robustness measures proposed. Results show that our proposed metrics reveal a clear trend of improved robustness to perturbations when subword regularization methods are used.

Gender in Danger? Evaluating Speech Translation Technology on the MuST-SHE Corpus [Website][PDF]
Luisa Bentivogli, Beatrice Savoldi, Matteo Negri, Mattia A. Di Gangi, Roldano Cattoni, and Marco Turchi 18:00–19:00

Translating from languages without productive grammatical gender like English into gender-marked languages is a well-known difficulty for machines. This difficulty is also due to the fact that the training data on which models are built typically reflect the asymmetries of natural languages, gender bias included. Exclusively fed with textual data, machine translation is intrinsically constrained by the fact that the input sentence does not always contain clues about the gender identity of the referred human entities. But what happens with speech translation, where the input is an audio signal? Can audio provide additional information to reduce gender bias? We present the first thorough investigation of gender bias in speech translation, contributing with: i) the release of a benchmark useful for future studies, and ii) the comparison of different technologies (cascade and end-to-end) on two language directions (English-Italian/French).

In Neural Machine Translation, What Does Transfer Learning Transfer? [Website][PDF]
Alham Fikri Aji, Nikolay Bogoychev, Kenneth Heafield, and Rico Sennrich 18:00–19:00

Transfer learning improves quality for low-resource machine translation, but it is unclear what exactly it transfers. We perform several ablation studies that limit information transfer, then measure the quality impact across three language pairs to gain a black-box understanding of transfer learning. Word embeddings play an important role in transfer learning, particularly if they are properly aligned. Although transfer learning can be performed without embeddings, results are sub-optimal. In contrast, transferring only the embeddings but nothing else yields catastrophic results. We then investigate diagonal alignments with auto-encoders over real languages and randomly generated se-

quences, finding even randomly generated sequences as parents yield noticeable but smaller gains. Finally, transfer learning can eliminate the need for a warm-up phase when training transformer models in high resource language pairs.

Parallel Corpus Filtering via Pre-trained Language Models [Website][PDF]
Boliang Zhang, Ajay Nagesh, and Kevin Knight 18:00–19:00

Web-crawled data provides a good source of parallel corpora for training machine translation models. It is automatically obtained, but extremely noisy, and recent work shows that neural machine translation systems are more sensitive to noise than traditional statistical machine translation methods. In this paper, we propose a novel approach to filter out noisy sentence pairs from web-crawled corpora via pre-trained language models. We measure sentence parallelism by leveraging the multilingual capability of BERT and use the Generative Pre-training (GPT) language model as a domain filter to balance data domains. We evaluate the proposed method on the WMT 2018 Parallel Corpus Filtering shared task, and on our own web-crawled Japanese-Chinese parallel corpus. Our method significantly outperforms baselines and achieves a new state-of-the-art. In an unsupervised setting, our method achieves comparable performance to the top-1 supervised method. We also evaluate on a web-crawled Japanese-Chinese parallel corpus that we make publicly available.

Reducing Gender Bias in Neural Machine Translation as a Domain Adaptation Problem [Website][PDF]
Danielle Saunders and Bill Byrne 18:00–19:00

Training data for NLP tasks often exhibits gender bias in that fewer sentences refer to women than to men. In Neural Machine Translation (NMT) gender bias has been shown to reduce translation quality, particularly when the target language has grammatical gender. The recent WinoMT challenge set allows us to measure this effect directly (Stanovsky et al, 2019) Ideally we would reduce system bias by simply debiasing all data prior to training, but achieving this effectively is itself a challenge. Rather than attempt to create a 'balanced' dataset, we use transfer learning on a small set of trusted, gender-balanced examples. This approach gives strong and consistent improvements in gender debiasing with much less computational cost than training from scratch. A known pitfall of transfer learning on new domains is 'catastrophic forgetting', which we address at adaptation and inference time. During adaptation we show that Elastic Weight Consolidation allows a performance trade-off between general translation quality and bias reduction. At inference time we propose a lattice-rescoring scheme which outperforms all systems evaluated in Stanovsky et al, 2019 on WinoMT with no degradation of general test set BLEU. We demonstrate our approach translating from English into three languages with varied linguistic properties and data availability.

Regularized Context Gates on Transformer for Machine Translation [Website][PDF]
Xintong Li, Lemao Liu, Rui Wang, Guoping Huang, and Max Meng 18:00–19:00

Context gates are effective to control the contributions from the source and target contexts in the recurrent neural network (RNN) based neural machine translation (NMT). However, it is challenging to extend them into the advanced Transformer architecture, which is more complicated than RNN. This paper first provides a method to identify source and target contexts and then introduce a gate mechanism to control the source and target contributions in Transformer. In addition, to further reduce the bias problem in the gate mechanism, this paper proposes a regularization method to guide the learning of the gates with supervision automatically generated using pointwise mutual information. Extensive experiments on 4 translation datasets demonstrate that the proposed model obtains an averaged gain of 1.0 BLEU score over a strong Transformer baseline.

Unsupervised Domain Clusters in Pretrained Language Models [Website][PDF]
Roee Aharoni and Yoav Goldberg 18:00–19:00

The notion of "in-domain data" in NLP is often over-simplistic and vague, as textual data varies in many nuanced linguistic aspects such as topic, style or level of formality. In addition, domain labels are many times unavailable, making it challenging to build domain-specific systems. We show that massive pre-trained language models implicitly learn sentence representations that cluster by domains without supervision – suggesting a simple data-driven definition of domains in textual data. We harness this property and propose domain data selection methods based on such models, which require only a small set of in-domain monolingual data. We evaluate our data selection methods for neural machine translation across five diverse domains, where they outperform an established approach as measured by both BLEU and precision and recall with respect to an oracle selection.

Using Context in Neural Machine Translation Training Objectives [Website][PDF]
Danielle Saunders, Felix Stahlberg, and Bill Byrne 18:00–19:00

We present Neural Machine Translation (NMT) training using document-level metrics with batch-level documents. Previous sequence-objective approaches to NMT training focus exclusively on sentence-level metrics like sentence BLEU which do not correspond to the desired evaluation metric, typically document BLEU. Meanwhile research into document-level NMT training focuses on data or model architecture rather than training procedure. We find that each of these lines of research has a clear space in it for the other, and propose merging them with a scheme that allows a document-level evaluation metric to be used in the NMT training objective. We first sample pseudo-documents from sentence samples. We then approximate the expected document BLEU gradient with Monte Carlo sampling for use as a cost function in Minimum Risk Training (MRT). This two-level sampling procedure gives NMT performance gains over sequence MRT and maximum-likelihood training. We demonstrate that training is more robust for document-level metrics than with sequence metrics. We further demonstrate improvements on NMT with TER and Grammatical Error Correction (GEC) using GLEU, both metrics used at the document level for evaluations.

Variational Neural Machine Translation with Normalizing Flows [Website][PDF]
Hendra Setiawan, Matthias Sperber, Udhyakumar Nallasamy, and Matthias Paulik 18:00–19:00

Variational Neural Machine Translation (VNMT) is an attractive framework for modeling the generation of target translations, conditioned not only on the source sentence but also on some latent random variables. The latent variable modeling may introduce useful statistical dependencies that can improve translation accuracy. Unfortunately, learning informative latent variables is non-trivial, as the latent space can be prohibitively large, and the latent codes are prone to be ignored by many translation models at training time. Previous works impose strong assumptions on the distribution of the latent code and limit the choice of the NMT architecture. In this paper, we propose to apply the VNMT framework to the state-of-the-art Transformer and introduce a more flexible approximate posterior based on normalizing flows. We demonstrate the efficacy of our proposal under both in-domain and out-of-domain conditions, significantly outperforming strong baselines.

Session 14B: NLP Applications-11

A Multi-Perspective Architecture for Semantic Code Search [Website][PDF]
Rajarshi Haldar, Lingfei Wu, JinJun Xiong, and Julia Hockenmaier 18:00–19:00

The ability to match pieces of code to their corresponding natural language descriptions and vice versa is fundamental for natural language search interfaces to software repositories. In this paper, we propose a novel multi-perspective cross-lingual neural framework for code–text matching, inspired in part by a previous model for monolingual text-to-text matching, to capture both global and local similarities. Our experiments on the CoNaLa dataset show that our proposed model yields better performance on this cross-lingual text-to-code matching task than previous approaches that map code and text to a single joint embedding space.

Automated Topical Component Extraction Using Neural Network Attention Scores from Source-based Essay Scoring [Website][PDF]
Haoran Zhang and Diane Litman 18:00–19:00

While automated essay scoring (AES) can reliably grade essays at scale, automated writing evaluation (AWE) additionally provides formative feedback to guide essay revision. However, a neural AES typically does not provide useful feature representations for supporting AWE. This paper presents a method for linking AWE and neural AES, by extracting Topical Components (TCs) representing evidence from a source text using the intermediate output of attention layers. We evaluate performance using a feature-based AES requiring TCs. Results show that performance is comparable whether using automatically or manually constructed TCs for 1) representing essays as rubric-based features, 2) grading essays.

Clinical Concept Linking with Contextualized Neural Representations [Website][PDF]
Elliot Schumacher, Andriy Mulyar, and Mark Dredze 18:00–19:00

In traditional approaches to entity linking, linking decisions are based on three sources of information – the similarity of the mention string to an entity's name, the similarity of the context of the document to the entity, and broader information about the knowledge base (KB). In some domains, there is little contextual information present in the KB and thus we rely more heavily on mention string similarity. We consider one example of this, concept linking, which seeks to link mentions of medical concepts to a medical concept ontology. We propose an approach to concept linking that leverages recent work in contextualized neural models, such as ELMo (Peters et al. 2018), which create a token representation that integrates the surrounding context of the mention and concept name. We find a neural ranking approach paired with contextualized embeddings provides gains over a competitive baseline (Leaman et al. 2013). Additionally, we find that a pre-training step using synonyms from the ontology offers a useful initialization for the ranker.

DeSePtion: Dual Sequence Prediction and Adversarial Examples for Improved Fact-Checking [Website][PDF]
Christopher Hidey, Tuhin Chakrabarty, Tariq Alhindi, Siddharth Varia, Kriste Krstovski, Mona Diab, and Smaranda Muresan 18:00–19:00

The increased focus on misinformation has spurred development of data and systems for detecting the veracity of a claim as well as retrieving authoritative evidence. The Fact Extraction and VERification (FEVER) dataset provides such a resource for evaluating endto- end fact-checking, requiring retrieval of evidence from Wikipedia to validate a veracity prediction. We show that current systems for FEVER are vulnerable to three categories of realistic challenges for fact-checking — multiple propositions, temporal reasoning, and ambiguity and lexical variation — and introduce a resource with these types of claims. Then we present a system designed to be resilient to these "attacks" using multiple pointer networks for document selection and jointly modeling a sequence of evidence sentences and veracity relation predictions. We find that in handling these attacks we obtain state-of-the-art results on FEVER, largely due to improved evidence retrieval.

Estimating predictive uncertainty for rumour verification models [Website][PDF]
Elena Kochkina and Maria Liakata 18:00–19:00

The inability to correctly resolve rumours circulating online can have harmful real-world consequences. We present a method for incorporating model and data uncertainty estimates into natural language processing models for automatic rumour verification. We show that these estimates can be used to filter out model predictions likely to be erroneous so that these difficult instances can be prioritised by a human fact-checker. We propose two methods for uncertainty-based instance rejection, supervised and unsupervised. We also show how uncertainty estimates can be used to interpret model performance as a rumour unfolds.

Let Me Choose: From Verbal Context to Font Selection [Website][PDF]
Amirreza Shirani, Franck Dernoncourt, Jose Echevarria, Paul Asente, Nedim Lipka, and Thamar Solorio
18:00–19:00

In this paper, we aim to learn associations between visual attributes of fonts and the verbal context of the texts they are typically applied to. Compared to related work leveraging the surrounding visual context, we choose to focus only on the input text, which can enable new applications for which the text is the only visual element in the document. We introduce a new dataset, containing examples of different topics in social media posts and ads, labeled through crowd-sourcing. Due to the subjective nature of the task, multiple fonts might be perceived as acceptable for an input text, which makes this problem challenging. To this end, we investigate different end-to-end models to learn label distributions on crowd-sourced data, to capture inter-subjectivity across all annotations.

[TACL] Machine Learning Driven Language Assessment [Website][PDF]
Burr Settles, Masato Hagiwara, and Geoffrey T. LaFlair 18:00–19:00

We describe a method for rapidly creating language proficiency assessments, and provide experimental evidence that such tests can be valid, reliable, and secure. Our approach is the first to use machine learning and natural language processing to induce proficiency scales based on a given standard, and then use linguistic models to estimate item difficulty directly for computer-adaptive testing. This alleviates the need for expensive pilot testing with human subjects. We used these methods to develop an online proficiency exam called the Duolingo English Test, and demonstrate that its scores align significantly with other high-stakes English assessments. Furthermore, our approach produces test scores that are highly reliable, while generating item banks large enough to satisfy security requirements.

Multi-Label and Multilingual News Framing Analysis [Website][PDF]
Afra Feyza Akyürek, Lei Guo, Randa Elanwar, Prakash Ishwar, Margrit Betke, and Derry Tanti Wijaya
18:00–19:00

News framing refers to the practice in which aspects of specific issues are highlighted in the news to promote a particular interpretation. In NLP, although recent works have studied framing in English news, few have studied how the analysis can be extended to other languages and in a multi-label setting. In this work, we explore multilingual transfer learning to detect multiple frames from just the news headline in a genuinely low-resource context where there are few/no frame annotations in the target language. We propose a novel method that can leverage elementary resources consisting of a dictionary and few annotations to detect frames in the target language. Our method performs comparably or better than translating the entire target language headline to the source language for which we have annotated data. This work opens up an exciting new capability of scaling up frame analysis to many languages, even those without existing translation technologies. Lastly, we apply our method to detect frames on the issue of U.S. gun violence in multiple languages and obtain exciting insights on the relationship between different frames of the same problem across different countries with different languages.

Predicting Performance for Natural Language Processing Tasks [Website][PDF]
Mengzhou Xia, Antonios Anastasopoulos, Ruochen Xu, Yiming Yang, and Graham Neubig 18:00–19:00

Given the complexity of combinations of tasks, languages, and domains in natural language processing (NLP) research, it is computationally prohibitive to exhaustively test newly proposed models on each possible experimental setting. In this work, we attempt to explore the possibility of gaining plausible judgments of how well an NLP model can perform under an experimental setting, *without actually training or testing the model*. To do so, we build regression models to predict the evaluation score of an NLP experiment given the experimental settings as input. Experimenting on–9 different NLP tasks, we find that our predictors can produce meaningful predictions over unseen languages and different modeling architectures, outperforming reasonable baselines as well as human experts. %we represent experimental settings using an array of features. Going further, we outline how our predictor can be used to find a small subset of representative experiments that should be run in order to obtain plausible predictions for all other experimental settings.[1]

ScriptWriter: Narrative-Guided Script Generation [Website][PDF]
Yutao Zhu, Ruihua Song, Zhicheng Dou, Jian-Yun NIE, and Jin Zhou 18:00–19:00

It is appealing to have a system that generates a story or scripts automatically from a storyline, even though this is still out of our reach. In dialogue systems, it would also be useful to drive dialogues by a dialogue plan. In this paper, we address a key problem involved in these applications - guiding a dialogue by a narrative. The proposed model ScriptWriter selects the best response among the candidates that fit the context as well as the given narrative. It keeps track of what in the narrative has been said and what is to be said. A narrative plays a different role than the context (i.e., previous utterances), which is generally used in current dialogue systems. Due to the unavailability of data for this new application, we construct a new large-scale data collection GraphMovie from a movie website where end-users can upload their narratives freely when watching a movie. Experimental results on the dataset show that our proposed approach based on narratives significantly outperforms the baselines that simply use the narrative as a kind of context.

Should All Cross-Lingual Embeddings Speak English? [Website][PDF]
Antonios Anastasopoulos and Graham Neubig 18:00–19:00

Most of recent work in cross-lingual word embeddings is severely Anglocentric. The vast majority of lexicon induction evaluation dictionaries are between English and another language, and the English embedding space is selected by default as the hub when learning in a multilingual setting. With this work, however, we challenge these practices. First, we show that the choice of hub language can significantly impact downstream lexicon induction zero-shot POS tagging performance. Second, we both expand a standard English-centered evaluation dictionary collection to include all language pairs using triangulation, and create new dictionaries for under-represented languages. Evaluating established methods over all these language pairs sheds light into their suitability for aligning embeddings from distant languages and presents new challenges for the field. Finally, in our analysis we identify general guidelines for strong cross-lingual embedding baselines, that extend to language pairs that do not include English.

Smart To-Do: Automatic Generation of To-Do Items from Emails [Website][PDF]
Sudipto Mukherjee, Subhabrata Mukherjee, Marcello Hasegawa, Ahmed Hassan Awadallah, and Ryen White 18:00–19:00

Intelligent features in email service applications aim to increase productivity by helping people organize their folders, compose their emails and respond to pending tasks. In this work, we explore a new application, Smart-To-Do, that

[1] Code, data and logs are publicly available at `https://github.com/xiamengzhou/NLPerf`.

helps users with task management over emails. We introduce a new task and dataset for automatically generating To-Do items from emails where the sender has promised to perform an action. We design a two-stage process leveraging recent advances in neural text generation and sequence-to-sequence learning, obtaining BLEU and ROUGE scores of 0.23 and 0.63 for this task. To the best of our knowledge, this is the first work to address the problem of composing To-Do items from emails.

Understanding Advertisements with BERT [Website] [PDF]

Kanika Kalra, Bhargav Kurma, Silpa Vadakkeeveetil Sreelatha, Manasi Patwardhan, and Shirish Karande 18:00–19:00

We consider a task based on CVPR 2018 challenge dataset on advertisement (Ad) understanding. The task involves detecting the viewer's interpretation of an Ad image captured as text. Recent results have shown that the embedded scene-text in the image holds a vital cue for this task. Motivated by this, we fine-tune the base BERT model for a sentence-pair classification task. Despite utilizing the scene-text as the only source of visual information, we could achieve a hit-or-miss accuracy of 84.95% on the challenge test data. To enable BERT to process other visual information, we append image captions to the scene-text. This achieves an accuracy of 89.69%, which is an improvement of 4.7%. This is the best reported result for this task.

Session 14B Semantics: Lexical-8

[CL] LESSLEX: Linking Multilingual Embeddings to SenSe Representations of Lexical Items [Website][PDF]
Davide Colla, Enrico Mensa, and Daniele P. Radicioni 18:00–19:00

We present LESSLEX, a novel multilingual lexical resource. Different from the vast majority of existing approaches, we ground our embeddings on a sense inventory made available from the BabelNet semantic network. In this setting, multilingual access is governed by the mapping of terms onto their underlying sense descriptions, such that all vectors co-exist in the same semantic space. As a result, for each term we have thus the 'blended' terminological vector along with those describing all senses associated to that term. LessLex has been tested on three tasks relevant to lexical semantics: conceptual similarity, contextual similarity, and semantic text similarity: we experimented over the principal data sets for such tasks in their multilingual and cross-lingual variants, improving on or closely approaching state-of-the-art results. We conclude by arguing that LessLex vectors may be relevant for practical applications and for research on conceptual and lexical access and competence.

Non-Linear Instance-Based Cross-Lingual Mapping for Non-Isomorphic Embedding Spaces [Website][PDF]
Goran Glavaš and Ivan Vulić 18:00–19:00

We present InstaMap, an instance-based method for learning projection-based cross-lingual word embeddings. Unlike prior work, it deviates from learning a single global linear projection. InstaMap is a non-parametric model that learns a non-linear projection by iteratively: (1) finding a globally optimal rotation of the source embedding space relying on the Kabsch algorithm, and then (2) moving each point along an instance-specific translation vector estimated from the translation vectors of the point's nearest neighbours in the training dictionary. We report performance gains with InstaMap over four representative state-of-the-art projection-based models on bilingual lexicon induction across a set of 28 diverse language pairs. We note prominent improvements, especially for more distant language pairs (i.e., languages with non-isomorphic monolingual spaces).

Session 14B Semantics: Textual Inference and Other Areas of Semantics-6

Are Natural Language Inference Models IMPPRESsive? Learning IMPlicature and PRESupposition [Website][PDF]
Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams 18:00–19:00

Natural language inference (NLI) is an increasingly important task for natural language understanding, which requires one to infer whether a sentence entails another. However, the ability of NLI models to make pragmatic inferences remains understudied. We create an IMPlicature and PRESupposition diagnostic dataset (IMPPRES), consisting of 32K semi-automatically generated sentence pairs illustrating well-studied pragmatic inference types. We use IMPPRES to evaluate whether BERT, InferSent, and BOW NLI models trained on MultiNLI (Williams et al., 2018) learn to make pragmatic inferences. Although MultiNLI appears to contain very few pairs illustrating these inference types, we find that BERT learns to draw pragmatic inferences. It reliably treats scalar implicatures triggered by "some" as entailments. For some presupposition triggers like "only", BERT reliably recognizes the presupposition as an entailment, even when the trigger is embedded under an entailment canceling operator like negation. BOW and InferSent show weaker evidence of pragmatic reasoning. We conclude that NLI training encourages models to learn some, but not all, pragmatic inferences.

Benchmarking Multimodal Regex Synthesis with Complex Structures [Website][PDF]
Xi Ye, Qiaochu Chen, Isil Dillig, and Greg Durrett 18:00–19:00

Existing datasets for regular expression (regex) generation from natural language are limited in complexity; compared to regex tasks that users post on StackOverflow, the regexes in these datasets are simple, and the language used to describe them is not diverse. We introduce StructuredRegex, a new regex synthesis dataset differing from prior ones in three aspects. First, to obtain structurally complex and realistic regexes, we generate the regexes using a probabilistic grammar with pre-defined macros observed from real-world StackOverflow posts. Second, to obtain linguistically diverse natural language descriptions, we show crowdworkers abstract depictions of the underlying regex and ask them to describe the pattern they see, rather than having them paraphrase synthetic language. Third, we augment each regex example with a collection of strings that are and are not matched by the ground truth regex, similar to how real users give examples. Our quantitative and qualitative analysis demonstrates the advantages of StructuredRegex over prior datasets. Further experimental results using various multimodal synthesis techniques highlight the challenge presented by our dataset, including non-local constraints and multi-modal inputs.

End-to-End Bias Mitigation by Modelling Biases in Corpora [Website][PDF]
Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson 18:00–19:00

Several recent studies have shown that strong natural language understanding (NLU) models are prone to relying on unwanted dataset biases without learning the underlying task, resulting in models that fail to generalize to out-of-domain datasets and are likely to perform poorly in real-world scenarios. We propose two learning strategies to train neural models, which are more robust to such biases and transfer better to out-of-domain datasets. The biases are specified in terms of one or more bias-only models, which learn to leverage the dataset biases. During training, the bias-only models' predictions are used to adjust the loss of the base model to reduce its reliance on biases by down-weighting the biased examples and focusing the training on the hard examples. We experiment on large-scale natural language inference and fact verification benchmarks, evaluating on out-of-domain datasets that are specifically designed to assess the robustness of models against known biases in the training data. Results show that our debiasing methods greatly improve robustness in all settings and better transfer to other textual entailment datasets. Our code and data are publicly available in `https://github.com/rabeehk/robust-nli`.

Generating Fact Checking Explanations [Website][PDF]
Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein 18:00–19:00

Most existing work on automated fact checking is concerned with predicting the veracity of claims based on metadata, social network spread, language used in claims, and, more recently, evidence supporting or denying claims. A crucial piece of the puzzle that is still missing is to understand how to automate the most elaborate part of the process – generating justifications for verdicts on claims. This paper provides the first study of how these explanations can be generated automatically based on available claim context, and how this task can be modelled jointly with veracity prediction. Our results indicate that optimising both objectives at the same time, rather than training them separately, improves the performance of a fact checking system. The results of a manual evaluation further suggest that the informativeness, coverage and overall quality of the generated explanations are also improved in the multi-task model.

Mind the Trade-off: Debiasing NLU Models without Degrading the In-distribution Performance [Website][PDF]
Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych 18:00–19:00

Models for natural language understanding (NLU) tasks often rely on the idiosyncratic biases of the dataset, which make them brittle against test cases outside the training distribution. Recently, several proposed debiasing methods are shown to be very effective in improving out-of-distribution performance. However, their improvements come at the expense of performance drop when models are evaluated on the in-distribution data, which contain examples with higher diversity. This seemingly inevitable trade-off may not tell us much about the changes in the reasoning and understanding capabilities of the resulting models on broader types of examples beyond the small subset represented in the out-of-distribution data. In this paper, we address this trade-off by introducing a novel debiasing method, called confidence regularization, which discourage models from exploiting biases while enabling them to receive enough incentive to learn from all the training examples. We evaluate our method on three NLU tasks and show that,

in contrast to its predecessors, it improves the performance on out-of-distribution datasets (e.g., 7pp gain on HANS dataset) while maintaining the original in-distribution accuracy.

NILE : Natural Language Inference with Faithful Natural Language Explanations [Website][PDF]
Sawan Kumar and Partha Talukdar 18:00–19:00

The recent growth in the popularity and success of deep learning models on NLP classification tasks has accompanied the need for generating some form of natural language explanation of the predicted labels. Such generated natural language (NL) explanations are expected to be faithful, i.e., they should correlate well with the model's internal decision making. In this work, we focus on the task of natural language inference (NLI) and address the following question: can we build NLI systems which produce labels with high accuracy, while also generating faithful explanations of its decisions? We propose Natural-language Inference over Label-specific Explanations (NILE), a novel NLI method which utilizes auto-generated label-specific NL explanations to produce labels along with its faithful explanation. We demonstrate NILE's effectiveness over previously reported methods through automated and human evaluation of the produced labels and explanations. Our evaluation of NILE also supports the claim that accurate systems capable of providing testable explanations of their decisions can be designed. We discuss the faithfulness of NILE's explanations in terms of sensitivity of the decisions to the corresponding explanations. We argue that explicit evaluation of faithfulness, in addition to label and explanation accuracy, is an important step in evaluating model's explanations. Further, we demonstrate that task-specific probes are necessary to establish such sensitivity.

Neural Mixed Counting Models for Dispersed Topic Discovery [Website][PDF]
Jiemin Wu, Yanghui Rao, Zusheng Zhang, Haoran Xie, Qing Li, Fu Lee Wang, and Ziye Chen 18:00–19:00

Mixed counting models that use the negative binomial distribution as the prior can well model over-dispersed and hierarchically dependent random variables; thus they have attracted much attention in mining dispersed document topics. However, the existing parameter inference method like Monte Carlo sampling is quite time-consuming. In this paper, we propose two efficient neural mixed counting models, i.e., the Negative Binomial-Neural Topic Model (NB-NTM) and the Gamma Negative Binomial-Neural Topic Model (GNB-NTM) for dispersed topic discovery. Neural variational inference algorithms are developed to infer model parameters by using the reparameterization of Gamma distribution and the Gaussian approximation of Poisson distribution. Experiments on real-world datasets indicate that our models outperform state-of-the-art baseline models in terms of perplexity and topic coherence. The results also validate that both NB-NTM and GNB-NTM can produce explainable intermediate variables by generating dispersed proportions of document topics.

QuASE: Question-Answer Driven Sentence Encoding [Website][PDF]
Hangfeng He, Qiang Ning, and Dan Roth 18:00–19:00

Question-answering (QA) data often encodes essential information in many facets. This paper studies a natural question: Can we get supervision from QA data for other tasks (typically, non-QA ones)? For example, *can we use QAMR (Michael et al., 2017) to improve named entity recognition?* We suggest that simply further pre-training BERT is often not the best option, and propose the *question-answer driven sentence encoding (QuASE)* framework. QuASE learns representations from QA data, using BERT or other state-of-the-art contextual language models. In particular, we observe the need to distinguish between two types of sentence encodings, depending on whether the target task is a single- or multi-sentence input; in both cases, the resulting encoding is shown to be an easy-to-use plugin for many downstream tasks. This work may point out an alternative way to supervise NLP tasks.

Temporal Common Sense Acquisition with Minimal Supervision [Website][PDF]
Ben Zhou, Qiang Ning, Daniel Khashabi, and Dan Roth 18:00–19:00

Temporal common sense (e.g., duration and frequency of events) is crucial for understanding natural language. However, its acquisition is challenging, partly because such information is often not expressed explicitly in text, and human annotation on such concepts is costly. This work proposes a novel sequence modeling approach that exploits explicit and implicit mentions of temporal common sense, extracted from a large corpus, to build TacoLM, a temporal common sense language model. Our method is shown to give quality predictions of various dimensions of temporal common sense (on UDST and a newly collected dataset from RealNews). It also produces representations of events for relevant tasks such as duration comparison, parent-child relations, event coreference and temporal QA (on TimeBank, HiEVE and MCTACO) that are better than using the standard BERT. Thus, it will be an important component of temporal NLP.

The Sensitivity of Language Models and Humans to Winograd Schema Perturbations [Website][PDF]
Mostafa Abdou, Vinit Ravishankar, Maria Barrett, Yonatan Belinkov, Desmond Elliott, and Anders Søgaard 18:00–19:00

Large-scale pretrained language models are the major driving force behind recent improvements in perfromance on the Winograd Schema Challenge, a widely employed test of commonsense reasoning ability. We show, however, with a new diagnostic dataset, that these models are sensitive to linguistic perturbations of the Winograd examples that minimally affect human understanding. Our results highlight interesting differences between humans and language models: language models are more sensitive to number or gender alternations and synonym replacements than humans, and humans are more stable and consistent in their predictions, maintain a much higher absolute performance, and perform better on non-associative instances than associative ones.

Towards Robustifying NLI Models Against Lexical Dataset Biases [Website][PDF]
Xiang Zhou and Mohit Bansal 18:00–19:00

While deep learning models are making fast progress on the task of Natural Language Inference, recent studies have also shown that these models achieve high accuracy by exploiting several dataset biases, and without deep under-

standing of the language semantics. Using contradiction-word bias and word-overlapping bias as our two bias examples, this paper explores both data-level and model-level debiasing methods to robustify models against lexical dataset biases. First, we debias the dataset through data augmentation and enhancement, but show that the model bias cannot be fully removed via this method. Next, we also compare two ways of directly debiasing the model without knowing what the dataset biases are in advance. The first approach aims to remove the label bias at the embedding level. The second approach employs a bag-of-words sub-model to capture the features that are likely to exploit the bias and prevents the original model from learning these biased features by forcing orthogonality between these two sub-models. We performed evaluations on new balanced datasets extracted from the original MNLI dataset as well as the NLI stress tests, and show that the orthogonality approach is better at debiasing the model while maintaining competitive overall accuracy.

Uncertain Natural Language Inference [Website][PDF]
Tongfei Chen, Zhengping Jiang, Adam Poliak, Keisuke Sakaguchi, and Benjamin Van Durme 18:00–19:00

We introduce Uncertain Natural Language Inference (UNLI), a refinement of Natural Language Inference (NLI) that shifts away from categorical labels, targeting instead the direct prediction of subjective probability assessments. We demonstrate the feasibility of collecting annotations for UNLI by relabeling a portion of the SNLI dataset under a probabilistic scale, where items even with the same categorical label differ in how likely people judge them to be true given a premise. We describe a direct scalar regression modeling approach, and find that existing categorically-labeled NLI data can be used in pre-training. Our best models correlate well with humans, demonstrating models are capable of more subtle inferences than the categorical bin assignment employed in current NLI tasks.

Session 14B: Student Research Workshop

Why is penguin more similar to polar bear than to sea gull? Analyzing conceptual knowledge in distributional models [Website][PDF]
Pia Sommerauer 18:00–19:00

What do powerful models of word mean- ing created from distributional data (e.g. Word2vec (Mikolov et al., 2013) BERT (Devlin et al., 2019) and ELMO (Peters et al., 2018)) represent? What causes words to be similar in the semantic space? What type of information is lacking? This thesis proposal presents a framework for investigating the information encoded in distributional semantic models. Several analysis methods have been suggested, but they have been shown to be limited and are not well understood. This approach pairs observations made on actual corpora with insights obtained from data manipulation experiments. The expected outcome is a better understanding of (1) the semantic information we can infer purely based on linguistic co-occurrence patterns and (2) the potential of distributional semantic models to pick up linguistic evidence.

Pointwise Paraphrase Appraisal is Potentially Problematic [Website][PDF]
Hannah Chen, Yangfeng Ji, and David Evans 18:00–19:00

The prevailing approach for training and evaluating paraphrase identification models is constructed as a binary classification problem: the model is given a pair of sentences, and is judged by how accurately it classifies pairs as either paraphrases or non-paraphrases. This pointwise-based evaluation method does not match well the objective of most real world applications, so the goal of our work is to understand how models which perform well under pointwise evaluation may fail in practice and find better methods for evaluating paraphrase identification models. As a first step towards that goal, we show that although the standard way of fine-tuning BERT for paraphrase identification by pairing two sentences as one sequence results in a model with state-of-the-art performance, that model may perform poorly on simple tasks like identifying pairs with two identical sentences. Moreover, we show that these models may even predict a pair of randomly-selected sentences with higher paraphrase score than a pair of identical ones.

A Geometry-Inspired Attack for Generating Natural Language Adversarial Examples [Website]
Zhao Meng and Roger Wattenhofer 18:00–19:00

Generating adversarial examples for natural language is hard, as natural language consists of discrete symbols and examples are often of variable lengths. In this paper, we propose a geometry-inspired attack for generating natural language adversarial examples. Our attack generates adversarial examples by iteratively approximating the decision boundary of deep neural networks. Experiments on two datasets with two different models show that our attack fools the models with high success rates, while only replacing a few words. Human evaluation shows that adversarial examples generated by our attack are hard for humans to recognize. Further experiments show that adversarial training can improve model robustness against our attack.

Enhancing Word Embeddings with Knowledge Extracted from Lexical Resources [Website][PDF]
Magdalena Biesialska, Bardia Rafieian, and Marta R. Costa-jussà 18:00–19:00

In this work, we present an effective method for semantic specialization of word vector representations. To this end, we use traditional word embeddings and apply specialization methods to better capture semantic relations between words. In our approach, we leverage external knowledge from rich lexical resources such as BabelNet. We also show that our proposed post-specialization method based on an adversarial neural network with the Wasserstein distance allows to gain improvements over state-of-the-art methods on two tasks: word similarity and dialog state tracking.

Session 14B Syntax: Tagging, Chunking and Parsing-5

[TACL] Deep Contextualized Self-training for Low Resource Dependency Parsing [Website][PDF]
Guy Rotman and Roi Reichart 18:00–19:00

Neural dependency parsing has proven very effective, achieving state-of-the-art results on numerous domains and languages. Unfortunately, it requires large amounts of labeled data, that is costly and laborious to create. In this paper we propose a self-training algorithm that alleviates this annotation bottleneck by training a parser on its own output. Our Deep Contextualized Self-training (DCST) algorithm utilizes representation models trained on sequence labeling tasks that are derived from the parser's output when applied to unlabeled data and integrates these models with the base parser through a gating mechanism. We conduct experiments across multiple languages, both in low resource in-domain and in cross-domain setups and demonstrate that DCST substantially outperforms traditional self-training as well as recent semi-supervised training methods.

Extracting Headless MWEs from Dependency Parse Trees: Parsing, Tagging, and Joint Modeling Approaches [Website][PDF]
Tianze Shi and Lillian Lee 18:00–19:00

An interesting and frequent type of multi-word expression (MWE) is the headless MWE, for which there are no true internal syntactic dominance relations; examples include many named entities ("Wells Fargo") and dates ("July 5, 2020") as well as certain productive constructions ("blow for blow", "day after day"). Despite their special status and prevalence, current dependency-annotation schemes require treating such flat structures as if they had internal syntactic heads, and most current parsers handle them in the same fashion as headed constructions. Meanwhile, outside the context of parsing, taggers are typically used for identifying MWEs, but taggers might benefit from structural information. We empirically compare these two common strategies—parsing and tagging—for predicting flat MWEs. Additionally, we propose an efficient joint decoding algorithm that combines scores from both strategies. Experimental results on the MWE-Aware English Dependency Corpus and on six non-English dependency treebanks with frequent flat structures show that: (1) tagging is more accurate than parsing for identifying flat-structure MWEs, (2) our joint decoder reconciles the two different views and, for non-BERT features, leads to higher accuracies, and (3) most of the gains result from feature sharing between the parsers and taggers.

Revisiting Higher-Order Dependency Parsers [Website][PDF]
Erick Fonseca and André F. T. Martins 18:00–19:00

Neural encoders have allowed dependency parsers to shift from higher-order structured models to simpler first-order ones, making decoding faster and still achieving better accuracy than non-neural parsers. This has led to a belief that neural encoders can implicitly encode structural constraints, such as siblings and grandparents in a tree. We tested this hypothesis and found that neural parsers may benefit from higher-order features, even when employing a powerful pre-trained encoder, such as BERT. While the gains of higher-order features are small in the presence of a powerful encoder, they are consistent for long-range dependencies and long sentences. In particular, higher-order models are more accurate on full sentence parses and on the exact match of modifier lists, indicating that they deal better with larger, more complex structures.

SeqVAT: Virtual Adversarial Training for Semi-Supervised Sequence Labeling [Website][PDF]
Luoxin Chen, Weitong Ruan, Xinyue Liu, and Jianhua Lu 18:00–19:00

Virtual adversarial training (VAT) is a powerful technique to improve model robustness in both supervised and semi-supervised settings. It is effective and can be easily adopted on lots of image classification and text classification tasks. However, its benefits to sequence labeling tasks such as named entity recognition (NER) have not been shown as significant, mostly, because the previous approach can not combine VAT with the conditional random field (CRF). CRF can significantly boost accuracy for sequence models by putting constraints on label transitions, which makes it an essential component in most state-of-the-art sequence labeling model architectures. In this paper, we propose SeqVAT, a method which naturally applies VAT to sequence labeling models with CRF. Empirical studies show that SeqVAT not only significantly improves the sequence labeling performance over baselines under supervised settings, but also outperforms state-of-the-art approaches under semi-supervised settings.

Tetra-Tagging: Word-Synchronous Parsing with Linear-Time Inference [Website][PDF]
Nikita Kitaev and Dan Klein 18:00–19:00

We present a constituency parsing algorithm that, like a supertagger, works by assigning labels to each word in a sentence. In order to maximally leverage current neural architectures, the model scores each word's tags in parallel, with minimal task-specific structure. After scoring, a left-to-right reconciliation phase extracts a tree in (empirically) linear time. Our parser achieves 95.4 F1 on the WSJ test set while also achieving substantial speedups compared to current state-of-the-art parsers with comparable accuracies.

Treebank Embedding Vectors for Out-of-domain Dependency Parsing [Website][PDF]
Joachim Wagner, James Barry, and Jennifer Foster 18:00–19:00

A recent advance in monolingual dependency parsing is the idea of a treebank embedding vector, which allows all treebanks for a particular language to be used as training data while at the same time allowing the model to prefer training data from one treebank over others and to select the preferred treebank at test time. We build on this idea by 1) introducing a method to predict a treebank vector for sentences that do not come from a treebank used in training, and 2) exploring what happens when we move away from predefined treebank embedding vectors during test time and instead devise tailored interpolations. We show that 1) there are interpolated vectors that are superior to the predefined ones, and 2) treebank vectors can be predicted with sufficient accuracy, for nine out of ten test languages,

to match the performance of an oracle approach that knows the most suitable predefined treebank embedding for the test set.

Demo Session 4C

Time: *18:30–19:15*

MMPE: A Multi-Modal Interface using Handwriting, Touch Reordering, and Speech Commands for Post-Editing Machine Translation [Website] [PDF]

Nico Herbig, Santanu Pal, Tim Düwel, Kalliopi Meladaki, Mahsa Monshizadeh, Vladislav Hnatovskiy, Antonio Krüger, and Josef van Genabith

The shift from traditional translation to post-editing (PE) of machine-translated (MT) text can save time and reduce errors, but it also affects the design of translation interfaces, as the task changes from mainly generating text to correcting errors within otherwise helpful translation proposals. Since this paradigm shift offers potential for modalities other than mouse and keyboard, we present MMPE, the first prototype to combine traditional input modes with pen, touch, and speech modalities for PE of MT. Users can directly cross out or hand-write new text, drag and drop words for reordering, or use spoken commands to update the text in place. All text manipulations are logged in an easily interpretable format to simplify subsequent translation process research. The results of an evaluation with professional translators suggest that pen and touch interaction are suitable for deletion and reordering tasks, while speech and multi-modal combinations of select & speech are considered suitable for replacements and insertions. Overall, experiment participants were enthusiastic about the new modalities and saw them as useful extensions to mouse & keyboard, but not as a complete substitute.

What's The Latest? A Question-driven News Chatbot [Website] [PDF]

Philippe Laban, John Canny, and Marti A. Hearst

This work describes an automatic news chatbot that draws content from a diverse set of news articles and creates conversations with a user about the news. Key components of the system include the automatic organization of news articles into topical chatrooms, integration of automatically generated questions into the conversation, and a novel method for choosing which questions to present which avoids repetitive suggestions. We describe the algorithmic framework and present the results of a usability study that shows that news readers using the system successfully engage in multi-turn conversations about specific news stories.

Demo Session 5A

Time: *20:00–20:45*

MixingBoard: a Knowledgeable Stylized Integrated Text Generation Platform [Website][PDF]
Xiang Gao, Michel Galley, and Bill Dolan

We present MixingBoard, a platform for quickly building demos with a focus on knowledge grounded stylized text generation. We unify existing text generation algorithms in a shared codebase and further adapt earlier algorithms for constrained generation. To borrow advantages from different models, we implement strategies for cross-model integration, from the token probability level to the latent space level. An interface to external knowledge is provided via a module that retrieves, on-the-fly, relevant knowledge from passages on the web or a document collection. A user interface for local development, remote webpage access, and a RESTful API are provided to make it simple for users to build their own demos.

DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation [Website][PDF]
Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan

We present a large, tunable neural conversational response generation model, DIALOGPT (dialogue generative pre-trained transformer). Trained on 147M conversation-like exchanges extracted from Reddit comment chains over a period spanning from 2005 through 2017, DialoGPT extends the Hugging Face PyTorch transformer to attain a performance close to human both in terms of automatic and human evaluation in single-turn dialogue settings. We show that conversational systems that leverage DialoGPT generate more relevant, contentful and context-consistent responses than strong baseline systems. The pre-trained model and training pipeline are publicly released to facilitate research into neural response generation and the development of more intelligent open-domain dialogue systems.

SUPP.AI: finding evidence for supplement-drug interactions [Website][PDF]
Lucy Wang, Oyvind Tafjord, Arman Cohan, Sarthak Jain, Sam Skjonsberg, Carissa Schoenick, Nick Botner, and Waleed Ammar

Dietary supplements are used by a large portion of the population, but information on their pharmacologic interactions is incomplete. To address this challenge, we present SUPP.AI, an application for browsing evidence of supplement-drug interactions (SDIs) extracted from the biomedical literature. We train a model to automatically extract supplement information and identify such interactions from the scientific literature. To address the lack of labeled data for SDI identification, we use labels of the closely related task of identifying drug-drug interactions (DDIs) for supervision. We fine-tune the contextualized word representations of the RoBERTa language model using labeled DDI data, and apply the fine-tuned model to identify supplement interactions. We extract 195k evidence sentences from 22M articles (P=0.82, R=0.58, F1=0.68) for 60k interactions. We create the SUPP.AI application for users to search evidence sentences extracted by our model. SUPP.AI is an attempt to close the information gap on dietary supplements by making up-to-date evidence on SDIs more discoverable for researchers, clinicians, and consumers. An informational video on how to use SUPP.AI is available at: https://youtu.be/dR0ucKdORwc

Session 15A Overview – Wednesday, July 8, 2020 20:00–21:00

Track A *Generation-14* Abstracts	[TACL] A Knowledge-Enhanced Pretraining Model for Commonsense Story Generation *Guan, Huang, Huang, Zhao, and Zhu* [Website][PDF]	BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension *Lewis, Liu, Goyal, Ghazvininejad, Mohamed, Levy, Stoyanov, and Zettlemoyer* [Website][PDF]	BLEURT: Learning Robust Metrics for Text Generation *Sellam, Das, and Parikh* [Website][PDF]	Distilling Knowledge Learned in BERT for Text Generation *Chen, Gan, Cheng, Liu, and Liu* [Website][PDF]	Improving Image Captioning with Better Use of Caption *Shi, Zhou, Qiu, and Zhu* [Website][PDF]
	Iterative Edit-Based Unsupervised Sentence Simplification *Kumar, Mou, Golab, and Vechtomova* [Website][PDF]	Neural CRF Model for Sentence Alignment in Text Simplification *Jiang, Maddela, Lan, Zhong, and Xu* [Website][PDF]	One Size Does Not Fit All: Generating and Evaluating Variable Number of Keyphrases *Yuan, Wang, Meng, Thaker, Brusilovsky, He, and Trischler* [Website][PDF]		
Track B *Information Extraction-11* Abstracts	A Generate-and-Rank Framework with Semantic Type Regularization for Biomedical Concept Normalization *Xu, Zhang, and Bethard* [Website][PDF]	A Joint Neural Model for Information Extraction with Global Features *Lin, Ji, Huang, and Wu* [Website][PDF]	A Two-Step Approach for Implicit Event Argument Detection *Zhang, Kong, Liu, Ma, and Hovy* [Website][PDF]	Document-Level Event Role Filler Extraction using Multi-Granularity Contextualized Encoding *Du and Cardie* [Website][PDF]	Learning Interpretable Relationships between Entities, Relations and Concepts via Bayesian Structure Learning on Open Domain Facts *Zhang, Sun, Feng, and Li* [Website][PDF]
	Multi-Sentence Argument Linking *Ebner, Xia, Culkin, Rawlins, and Van Durme* [Website][PDF]	Revisiting Unsupervised Relation Extraction *Tran, Le, and Ananiadou* [Website][PDF]	Temporally-Informed Analysis of Named Entity Recognition *Rijhwani and Preotiuc-Pietro* [Website][PDF]	Towards Open Domain Event Trigger Identification using Adversarial Domain Adaptation *Naik and Rose* [Website][PDF]	ZeroShotCeres: Zero-Shot Relation Extraction from Semi-Structured Webpages *Lockard, Shiralkar, Dong, and Hajishirzi* [Website][PDF]
Track C *Information Retrieval and Text Mining-8* Abstracts	A Prioritization Model for Suicidality Risk Assessment *Shing, Resnik, and Oard* [Website][PDF]	CluHTM - Semantic Hierarchical Topic Modeling based on CluWords *Viegas, Cunha, Gomes, Pereira, Rocha, and Goncalves* [Website][PDF]	Empower Entity Set Expansion via Language Model Probing *Zhang, Shen, Shang, and Han* [Website][PDF]	Feature Projection for Improved Text Classification *Qin, Hu, and Liu* [Website][PDF]	Learning Robust Models for e-Commerce Product Search *Nguyen, Rao, and Subbian* [Website][PDF]
Track D *Language Grounding to Vision, Robotics and Beyond-9* Abstracts	A negative case analysis of visual grounding methods for VQA *Shrestha, Kafle, and Kanan* [Website][PDF]	Cross-Modality Relevance for Reasoning on Language and Vision *Zheng, Guo, and Kordjamshidi* [Website][PDF]	History for Visual Dialog: Do we really need it? *Agarwal, Bui, Lee, Konstas, and Rieser* [Website][PDF]	Knowledge Supports Visual Language Grounding: A Case Study on Colour Terms *Schüz and Zarrieß* [Website][PDF]	Learning Web-based Procedures by Reasoning over Explanations and Demonstrations in Context *Srivastava, Polozov, Jojic, and Meek* [Website][PDF]

	Mapping Natural Language Instructions to Mobile UI Action Sequences *Li, He, Zhou, Zhang, and Baldridge* [Website][PDF]	Refer360°: A Referring Expression Recognition Dataset in 360° Images *Cirik, Berg-Kirkpatrick, and Morency* [Website][PDF]	TVQA+: Spatio-Temporal Grounding for Video Question Answering *Lei, Yu, Berg, and Bansal* [Website][PDF]	Unsupervised Multimodal Neural Machine Translation with Pseudo Visual Pivoting *Huang, Hu, Chang, and Hauptmann* [Website][PDF]	Words Aren't Enough, Their Order Matters: On the Robustness of Grounding Visual Referring Expressions *Akula, Gella, Al-Onaizan, Zhu, and Reddy* [Website][PDF]
Track E *Machine Translation-17* Abstracts	Addressing Posterior Collapse with Mutual Information for Improved Variational Neural Machine Translation *McCarthy, Li, Gu, and Dong* [Website][PDF]	Evaluating Robustness to Input Perturbations for Neural Machine Translation *Niu, Mathur, Dinu, and Al-Onaizan* [Website][PDF]	Hard-Coded Gaussian Attention for Neural Machine Translation *You, Sun, and Iyyer* [Website][PDF]	Learning a Multi-Domain Curriculum for Neural Machine Translation *Wang, Tian, Ngiam, Yang, Caswell, and Parekh* [Website][PDF]	Tagged Back-translation Revisited: Why Does It Really Work? *Marie, Rubino, and Fujita* [Website][PDF]
Track F *Sentence Level-10* Abstracts	Active Learning for Coreference Resolution using Discrete Annotation *Li, Stanovsky, and Zettlemoyer* [Website][PDF]	[TACL] Decoding Brain Activity Associated with Literal and Metaphoric Sentence Comprehension using Distributional Semantic Models *Djokic, Maillard, Bulat, and Shutova* [Website][PDF]	Emerging Cross-lingual Structure in Pretrained Language Models *Conneau, Wu, Li, Zettlemoyer, and Stoyanov* [Website][PDF]	Estimating Mutual Information Between Dense Word Embeddings *Zhelezniak, Savkov, and Hammerla* [Website][PDF]	Exploring Unexplored Generalization Challenges for Cross-Database Semantic Parsing *Suhr, Chang, Shaw, and Lee* [Website][PDF]
	Good-Enough Compositional Data Augmentation *Andreas* [Website][PDF]	Incorporating External Knowledge through Pre-training for Natural Language to Code Generation *Xu, Jiang, Yin, Vasilescu, and Neubig* [Website][PDF]	Predicting the Focus of Negation: Model and Error Analysis *Hossain, Hamilton, Palmer, and Blanco* [Website][PDF]	TaBERT: Pre-training for Joint Understanding of Textual and Tabular Data *Yin, Neubig, Yih, and Riedel* [Website][PDF]	Unsupervised Cross-lingual Representation Learning at Scale *Conneau, Khandelwal, Goyal, Chaudhary, Wenzek, Guzmán, Grave, Ott, Zettlemoyer, and Stoyanov* [Website][PDF]
Track G *Textual Inference and Other Areas of Semantics-7* Abstracts	Are Natural Language Inference Models IMPPRESsive? Learning IMPlicature and PRESupposition *Jeretic, Warstadt, Bhooshan, and Williams* [Website][PDF]	End-to-End Bias Mitigation by Modelling Biases in Corpora *Karimi Mahabadi, Belinkov, and Henderson* [Website][PDF]	Mind the Trade-off: Debiasing NLU Models without Degrading the In-distribution Performance *Utama, Moosavi, and Gurevych* [Website][PDF]	NILE : Natural Language Inference with Faithful Natural Language Explanations *Kumar and Talukdar* [Website][PDF]	QuASE: Question-Answer Driven Sentence Encoding *He, Ning, and Roth* [Website][PDF]
	Towards Robustifying NLI Models Against Lexical Dataset Biases *Zhou and Bansal* [Website][PDF]	Uncertain Natural Language Inference *Chen, Jiang, Poliak, Sakaguchi, and Van Durme* [Website][PDF]			

Track H *Student Research Workshop* Abstracts	Checkpoint Reranking: An Approach to Select Better Hypothesis for Neural Machine Translation Systems *Pandramish and Sharma* [Website] [PDF]	Story-level Text Style Transfer: A Proposal *Qian* [Website] [PDF]	Non-Topical Coherence in Social Talk: A Call for Dialogue Model Enrichment *Luu and Malamud* [Website] [PDF]	Compositional Generalization by Factorizing Alignment and Translation *Russin, Jo, O'Reilly, and Bengio* [Website] [PDF]	

Session 15A Details

Session 15A: Generation-14

[TACL] A Knowledge-Enhanced Pretraining Model for Commonsense Story Generation [Website] [PDF]
Jian Guan, Fei Huang, Minlie Huang, Zhihao Zhao, and Xiaoyan Zhu 20:00–21:00

Story generation, namely generating a reasonable story from a leading context, is an important but challenging task. In spite of the success in modeling fluency and local coherence, existing neural language generation models (e.g., GPT-2) still suffer from repetition, logic conflicts, and lack of long-range coherence in generated stories. We conjecture that this is because of the difficulty of associating relevant commonsense knowledge, understanding the causal relationships, and planning entities and events with proper temporal order. In this paper, we devise a knowledge-enhanced pretraining model for commonsense story generation. We propose to utilize commonsense knowledge from external knowledge bases to generate reasonable stories. To further capture the causal and temporal dependencies between the sentences in a reasonable story, we employ multi-task learning which combines a discriminative objective to distinguish true and fake stories during fine-tuning. Automatic and manual evaluation shows that our model can generate more reasonable stories than state-of-the-art baselines, particularly in terms of logic and global coherence.

BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension [Website] [PDF]
Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer 20:00–21:00

We present BART, a denoising autoencoder for pretraining sequence-to-sequence models. BART is trained by (1) corrupting text with an arbitrary noising function, and (2) learning a model to reconstruct the original text. It uses a standard Tranformer-based neural machine translation architecture which, despite its simplicity, can be seen as generalizing BERT (due to the bidirectional encoder), GPT (with the left-to-right decoder), and other recent pretraining schemes. We evaluate a number of noising approaches, finding the best performance by both randomly shuffling the order of sentences and using a novel in-filling scheme, where spans of text are replaced with a single mask token. BART is particularly effective when fine tuned for text generation but also works well for comprehension tasks. It matches the performance of RoBERTa on GLUE and SQuAD, and achieves new state-of-the-art results on a range of abstractive dialogue, question answering, and summarization tasks, with gains of up to 3.5 ROUGE. BART also provides a 1.1 BLEU increase over a back-translation system for machine translation, with only target language pretraining. We also replicate other pretraining schemes within the BART framework, to understand their effect on end-task performance.

BLEURT: Learning Robust Metrics for Text Generation [Website] [PDF]
Thibault Sellam, Dipanjan Das, and Ankur Parikh 20:00–21:00

Text generation has made significant advances in the last few years. Yet, evaluation metrics have lagged behind, as the most popular choices (e.g., BLEU and ROUGE) may correlate poorly with human judgment. We propose BLEURT, a learned evaluation metric for English based on BERT. BLEURT can model human judgment with a few thousand possibly biased training examples. A key aspect of our approach is a novel pre-training scheme that uses millions of synthetic examples to help the model generalize. BLEURT provides state-of-the-art results on the last three years of the WMT Metrics shared task and the WebNLG data set. In contrast to a vanilla BERT-based approach, it yields superior results even when the training data is scarce and out-of-distribution.

Distilling Knowledge Learned in BERT for Text Generation [Website] [PDF]
Yen-Chun Chen, Zhe Gan, Yu Cheng, Jingzhou Liu, and Jingjing Liu 20:00–21:00

Large-scale pre-trained language model such as BERT has achieved great success in language understanding tasks. However, it remains an open question how to utilize BERT for language generation. In this paper, we present a novel approach, Conditional Masked Language Modeling (C-MLM), to enable the finetuning of BERT on target generation tasks. The finetuned BERT (teacher) is exploited as extra supervision to improve conventional Seq2Seq models (student) for better text generation performance. By leveraging BERT's idiosyncratic bidirectional nature, distilling knowledge learned in BERT can encourage auto-regressive Seq2Seq models to plan ahead, imposing global sequence-level supervision for coherent text generation. Experiments show that the proposed approach significantly outperforms strong Transformer baselines on multiple language generation tasks such as machine translation and text summarization. Our proposed model also achieves new state of the art on IWSLT German-English and English-Vietnamese MT datasets.

Improving Image Captioning with Better Use of Caption [Website] [PDF]
Zhan Shi, Xu Zhou, Xipeng Qiu, and Xiaodan Zhu 20:00–21:00

Image captioning is a multimodal problem that has drawn extensive attention in both the natural language processing and computer vision community. In this paper, we present a novel image captioning architecture to better explore semantics available in captions and leverage that to enhance both image representation and caption generation. Our models first construct caption-guided visual relationship graphs that introduce beneficial inductive bias using weakly supervised multi-instance learning. The representation is then enhanced with neighbouring and contextual nodes with their textual and visual features. During generation, the model further incorporates visual relationships using multi-task learning for jointly predicting word and object/predicate tag sequences. We perform extensive experiments on the MSCOCO dataset, showing that the proposed framework significantly outperforms the baselines, resulting in

the state-of-the-art performance under a wide range of evaluation metrics. The code of our paper has been made publicly available.

Iterative Edit-Based Unsupervised Sentence Simplification [Website][PDF]
Dhruv Kumar, Lili Mou, Lukasz Golab, and Olga Vechtomova 20:00–21:00

We present a novel iterative, edit-based approach to unsupervised sentence simplification. Our model is guided by a scoring function involving fluency, simplicity, and meaning preservation. Then, we iteratively perform word and phrase-level edits on the complex sentence. Compared with previous approaches, our model does not require a parallel training set, but is more controllable and interpretable. Experiments on Newsela and WikiLarge datasets show that our approach is nearly as effective as state-of-the-art supervised approaches.

Neural CRF Model for Sentence Alignment in Text Simplification [Website][PDF]
Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu 20:00–21:00

The success of a text simplification system heavily depends on the quality and quantity of complex-simple sentence pairs in the training corpus, which are extracted by aligning sentences between parallel articles. To evaluate and improve sentence alignment quality, we create two manually annotated sentence-aligned datasets from two commonly used text simplification corpora, Newsela and Wikipedia. We propose a novel neural CRF alignment model which not only leverages the sequential nature of sentences in parallel documents but also utilizes a neural sentence pair model to capture semantic similarity. Experiments demonstrate that our proposed approach outperforms all the previous work on monolingual sentence alignment task by more than 5 points in F1. We apply our CRF aligner to construct two new text simplification datasets, NEWSELA-AUTO and WIKI-AUTO, which are much larger and of better quality compared to the existing datasets. A Transformer-based seq2seq model trained on our datasets establishes a new state-of-the-art for text simplification in both automatic and human evaluation.

One Size Does Not Fit All: Generating and Evaluating Variable Number of Keyphrases [Website][PDF]
Xingdi Yuan, Tong Wang, Rui Meng, Khushboo Thaker, Peter Brusilovsky, Daqing He, and Adam Trischler 20:00–21:00

Different texts shall by nature correspond to different number of keyphrases. This desideratum is largely missing from existing neural keyphrase generation models. In this study, we address this problem from both modeling and evaluation perspectives. We first propose a recurrent generative model that generates multiple keyphrases as delimiter-separated sequences. Generation diversity is further enhanced with two novel techniques by manipulating decoder hidden states. In contrast to previous approaches, our model is capable of generating diverse keyphrases and controlling number of outputs. We further propose two evaluation metrics tailored towards the variable-number generation. We also introduce a new dataset StackEx that expands beyond the only existing genre (i.e., academic writing) in keyphrase generation tasks. With both previous and new evaluation metrics, our model outperforms strong baselines on all datasets.

Session 15A: Information Extraction-11

A Generate-and-Rank Framework with Semantic Type Regularization for Biomedical Concept Normalization [Website][PDF]
Dongfang Xu, Zeyu Zhang, and Steven Bethard 20:00–21:00

Concept normalization, the task of linking textual mentions of concepts to concepts in an ontology, is challenging because ontologies are large. In most cases, annotated datasets cover only a small sample of the concepts, yet concept normalizers are expected to predict all concepts in the ontology. In this paper, we propose an architecture consisting of a candidate generator and a list-wise ranker based on BERT. The ranker considers pairings of concept mentions and candidate concepts, allowing it to make predictions for any concept, not just those seen during training. We further enhance this list-wise approach with a semantic type regularizer that allows the model to incorporate semantic type information from the ontology during training. Our proposed concept normalization framework achieves state-of-the-art performance on multiple datasets.

A Joint Neural Model for Information Extraction with Global Features [Website][PDF]
Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu 20:00–21:00

Most existing joint neural models for Information Extraction (IE) use local task-specific classifiers to predict labels for individual instances (e.g., trigger, relation) regardless of their interactions. For example, a victim of a die event is likely to be a victim of an attack event in the same sentence. In order to capture such cross-subtask and cross-instance inter-dependencies, we propose a joint neural framework, OneIE, that aims to extract the globally optimal IE result as a graph from an input sentence. OneIE performs end-to-end IE in four stages: (1) Encoding a given sentence as contextualized word representations; (2) Identifying entity mentions and event triggers as nodes; (3) Computing label scores for all nodes and their pairwise links using local classifiers; (4) Searching for the globally optimal graph with a beam decoder. At the decoding stage, we incorporate global features to capture the cross-subtask and cross-instance interactions. Experiments show that adding global features improves the performance of our model and achieves new state of-the-art on all subtasks. In addition, as OneIE does not use any language-specific feature, we prove it can be easily applied to new languages or trained in a multilingual manner.

A Two-Step Approach for Implicit Event Argument Detection [Website][PDF]
Zhisong Zhang, Xiang Kong, Zhengzhong Liu, Xuezhe Ma, and Eduard Hovy 20:00–21:00

In this work, we explore the implicit event argument detection task, which studies event arguments beyond sentence boundaries. The addition of cross-sentence argument candidates imposes great challenges for modeling. To reduce the number of candidates, we adopt a two-step approach, decomposing the problem into two sub-problems: argument head-word detection and head-to-span expansion. Evaluated on the recent RAMS dataset (Ebner et al., 2020), our model achieves overall better performance than a strong sequence labeling baseline. We further provide detailed error analysis, presenting where the model mainly makes errors and indicating directions for future improvements. It remains a challenge to detect implicit arguments, calling for more future work of document-level modeling for this task.

Document-Level Event Role Filler Extraction using Multi-Granularity Contextualized Encoding [Website][PDF]
Xinya Du and Claire Cardie 20:00–21:00

Few works in the literature of event extraction have gone beyond individual sentences to make extraction decisions. This is problematic when the information needed to recognize an event argument is spread across multiple sentences. We argue that document-level event extraction is a difficult task since it requires a view of a larger context to determine which spans of text correspond to event role fillers. We first investigate how end-to-end neural sequence models (with pre-trained language model representations) perform on document-level role filler extraction, as well as how the length of context captured affects the models' performance. To dynamically aggregate information captured by neural representations learned at different levels of granularity (e.g., the sentence- and paragraph-level), we propose a novel multi-granularity reader. We evaluate our models on the MUC-4 event extraction dataset, and show that our best system performs substantially better than prior work. We also report findings on the relationship between context length and neural model performance on the task.

Learning Interpretable Relationships between Entities, Relations and Concepts via Bayesian Structure Learning on Open Domain Facts [Website][PDF]
Jingyuan Zhang, Mingming Sun, Yue Feng, and Ping Li 20:00–21:00

Concept graphs are created as universal taxonomies for text understanding in the open-domain knowledge. The nodes in concept graphs include both entities and concepts. The edges are from entities to concepts, showing that an entity is an instance of a concept. In this paper, we propose the task of learning interpretable relationships from open-domain facts to enrich and refine concept graphs. The Bayesian network structures are learned from open-domain facts as the interpretable relationships between relations of facts and concepts of entities. We conduct extensive experiments on public English and Chinese datasets. Compared to the state-of-the-art methods, the learned network structures help improving the identification of concepts for entities based on the relations of entities on both datasets.

Multi-Sentence Argument Linking [Website][PDF]
Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, and Benjamin Van Durme 20:00–21:00

We present a novel document-level model for finding argument spans that fill an event's roles, connecting related ideas in sentence-level semantic role labeling and coreference resolution. Because existing datasets for cross-sentence

linking are small, development of our neural model is supported through the creation of a new resource, Roles Across Multiple Sentences (RAMS), which contains 9,124 annotated events across 139 types. We demonstrate strong performance of our model on RAMS and other event-related datasets.

Revisiting Unsupervised Relation Extraction [Website][PDF]
Thy Thy Tran, Phong Le, and Sophia Ananiadou 20:00–21:00

Unsupervised relation extraction (URE) extracts relations between named entities from raw text without manually-labelled data and existing knowledge bases (KBs). URE methods can be categorised into generative and discriminative approaches, which rely either on hand-crafted features or surface form. However, we demonstrate that by using only named entities to induce relation types, we can outperform existing methods on two popular datasets. We conduct a comparison and evaluation of our findings with other URE techniques, to ascertain the important features in URE. We conclude that entity types provide a strong inductive bias for URE.

Temporally-Informed Analysis of Named Entity Recognition [Website][PDF]
Shruti Rijhwani and Daniel Preotiuc-Pietro 20:00–21:00

Natural language processing models often have to make predictions on text data that evolves over time as a result of changes in language use or the information described in the text. However, evaluation results on existing data sets are seldom reported by taking the timestamp of the document into account. We analyze and propose methods that make better use of temporally-diverse training data, with a focus on the task of named entity recognition. To support these experiments, we introduce a novel data set of English tweets annotated with named entities. We empirically demonstrate the effect of temporal drift on performance, and how the temporal information of documents can be used to obtain better models compared to those that disregard temporal information. Our analysis gives insights into why this information is useful, in the hope of informing potential avenues of improvement for named entity recognition as well as other NLP tasks under similar experimental setups.

Towards Open Domain Event Trigger Identification using Adversarial Domain Adaptation [Website][PDF]
Aakanksha Naik and Carolyn Rose 20:00–21:00

We tackle the task of building supervised event trigger identification models which can generalize better across domains. Our work leverages the adversarial domain adaptation (ADA) framework to introduce domain-invariance. ADA uses adversarial training to construct representations that are predictive for trigger identification, but not predictive of the example's domain. It requires no labeled data from the target domain, making it completely unsupervised. Experiments with two domains (English literature and news) show that ADA leads to an average F1 score improvement of 3.9 on out-of-domain data. Our best performing model (BERT-A) reaches 44-49 F1 across both domains, using no labeled target data. Preliminary experiments reveal that finetuning on 1% labeled data, followed by self-training leads to substantial improvement, reaching 51.5 and 67.2 F1 on literature and news respectively.

ZeroShotCeres: Zero-Shot Relation Extraction from Semi-Structured Webpages [Website][PDF]
Colin Lockard, Prashant Shiralkar, Xin Luna Dong, and Hannaneh Hajishirzi 20:00–21:00

In many documents, such as semi-structured webpages, textual semantics are augmented with additional information conveyed using visual elements including layout, font size, and color. Prior work on information extraction from semi-structured websites has required learning an extraction model specific to a given template via either manually labeled or distantly supervised data from that template. In this work, we propose a solution for "zero-shot" open-domain relation extraction from webpages with a previously unseen template, including from websites with little overlap with existing sources of knowledge for distant supervision and websites in entirely new subject verticals. Our model uses a graph neural network-based approach to build a rich representation of text fields on a webpage and the relationships between them, enabling generalization to new templates. Experiments show this approach provides a 31% F1 gain over a baseline for zero-shot extraction in a new subject vertical.

Session 15A: Information Retrieval and Text Mining-8

A Prioritization Model for Suicidality Risk Assessment [Website][PDF]
Han-Chin Shing, Philip Resnik, and Douglas Oard 20:00–21:00

We reframe suicide risk assessment from social media as a ranking problem whose goal is maximizing detection of severely at-risk individuals given the time available. Building on measures developed for resource-bounded document retrieval, we introduce a well founded evaluation paradigm, and demonstrate using an expert-annotated test collection that meaningful improvements over plausible cascade model baselines can be achieved using an approach that jointly ranks individuals and their social media posts.

CluHTM - Semantic Hierarchical Topic Modeling based on CluWords [Website][PDF]
Felipe Viegas, Washington Cunha, Christian Gomes, Antônio Pereira, Leonardo Rocha, and Marcos Goncalves 20:00–21:00

Hierarchical Topic modeling (HTM) exploits latent topics and relationships among them as a powerful tool for data analysis and exploration. Despite advantages over traditional topic modeling, HTM poses its own challenges, such as (1) topic incoherence, (2) unreasonable (hierarchical) structure, and (3) issues related to the definition of the "ideal" number of topics and depth of the hierarchy. In this paper, we advance the state-of-the-art on HTM by means of the design and evaluation of CluHTM, a novel non-probabilistic hierarchical matrix factorization aimed at solving the specific issues of HTM. CluHTM's novel contributions include: (i) the exploration of richer text representation that encapsulates both, global (dataset level) and local semantic information – when combined, these pieces of information help to solve the topic incoherence problem as well as issues related to the unreasonable structure; (ii) the exploitation of a stability analysis metric for defining the number of topics and the "shape" the hierarchical structure. In our evaluation, considering twelve datasets and seven state-of-the-art baselines, CluHTM outperformed the baselines in the vast majority of the cases, with gains of around \$500%\$ over the strongest state-of-the-art baselines. We also provide qualitative and quantitative statistical analyses of why our solution works so well.

Empower Entity Set Expansion via Language Model Probing [Website][PDF]
Yunyi Zhang, Jiaming Shen, Jingbo Shang, and Jiawei Han 20:00–21:00

Entity set expansion, aiming at expanding a small seed entity set with new entities belonging to the same semantic class, is a critical task that benefits many downstream NLP and IR applications, such as question answering, query understanding, and taxonomy construction. Existing set expansion methods bootstrap the seed entity set by adaptively selecting context features and extracting new entities. A key challenge for entity set expansion is to avoid selecting ambiguous context features which will shift the class semantics and lead to accumulative errors in later iterations. In this study, we propose a novel iterative set expansion framework that leverages automatically generated class names to address the semantic drift issue. In each iteration, we select one positive and several negative class names by probing a pre-trained language model, and further score each candidate entity based on selected class names. Experiments on two datasets show that our framework generates high-quality class names and outperforms previous state-of-the-art methods significantly.

Feature Projection for Improved Text Classification [Website][PDF]
Qi Qin, Wenpeng Hu, and Bing Liu 20:00–21:00

In classification, there are usually some good features that are indicative of class labels. For example, in sentiment classification, words like good and nice are indicative of the positive sentiment and words like bad and terrible are indicative of the negative sentiment. However, there are also many common features (e.g., words) that are not indicative of any specific class (e.g., voice and screen, which are common to both sentiment classes and are not discriminative for classification). Although deep learning has made significant progresses in generating discriminative features through its powerful representation learning, we believe there is still room for improvement. In this paper, we propose a novel angle to further improve this representation learning, i.e., feature projection. This method projects existing features into the orthogonal space of the common features. The resulting projection is thus perpendicular to the common features and more discriminative for classification. We apply this new method to improve CNN, RNN, Transformer, and Bert based text classification and obtain markedly better results.

Learning Robust Models for e-Commerce Product Search [Website][PDF]
Thanh Nguyen, Nikhil Rao, and Karthik Subbian 20:00–21:00

Showing items that do not match search query intent degrades customer experience in e-commerce. These mismatches result from counterfactual biases of the ranking algorithms toward noisy behavioral signals such as clicks and purchases in the search logs. Mitigating the problem requires a large labeled dataset, which is expensive and time-consuming to obtain. In this paper, we develop a deep, end-to-end model that learns to effectively classify mismatches and to generate hard mismatched examples to improve the classifier. We train the model end-to-end by introducing a latent variable into the cross-entropy loss that alternates between using the real and generated samples. This not only makes the classifier more robust but also boosts the overall ranking performance. Our model achieves a relative gain compared to baselines by over \$26%\$ in F-score, and over \$17%\$ in Area Under PR curve. On live search traffic, our model gains significant improvement in multiple countries.

Session 15A: Language Grounding to Vision, Robotics and Beyond-9

A negative case analysis of visual grounding methods for VQA [Website][PDF]
Robik Shrestha, Kushal Kafle, and Christopher Kanan 20:00–21:00

Existing Visual Question Answering (VQA) methods tend to exploit dataset biases and spurious statistical correlations, instead of producing right answers for the right reasons. To address this issue, recent bias mitigation methods for VQA propose to incorporate visual cues (e.g., human attention maps) to better ground the VQA models, showcasing impressive gains. However, we show that the performance improvements are not a result of improved visual grounding, but a regularization effect which prevents over-fitting to linguistic priors. For instance, we find that it is not actually necessary to provide proper, human-based cues; random, insensible cues also result in similar improvements. Based on this observation, we propose a simpler regularization scheme that does not require any external annotations and yet achieves near state-of-the-art performance on VQA-CPv2.

Cross-Modality Relevance for Reasoning on Language and Vision [Website][PDF]
Chen Zheng, Quan Guo, and Parisa Kordjamshidi 20:00–21:00

This work deals with the challenge of learning and reasoning over language and vision data for the related downstream tasks such as visual question answering (VQA) and natural language for visual reasoning (NLVR). We design a novel cross-modality relevance module that is used in an end-to-end framework to learn the relevance representation between components of various input modalities under the supervision of a target task, which is more generalizable to unobserved data compared to merely reshaping the original representation space. In addition to modeling the relevance between the textual entities and visual entities, we model the higher-order relevance between entity relations in the text and object relations in the image. Our proposed approach shows competitive performance on two different language and vision tasks using public benchmarks and improves the state-of-the-art published results. The learned alignments of input spaces and their relevance representations by NLVR task boost the training efficiency of VQA task.

History for Visual Dialog: Do we really need it? [Website][PDF]
Shubham Agarwal, Trung Bui, Joon-Young Lee, Ioannis Konstas, and Verena Rieser 20:00–21:00

Visual Dialogue involves "understanding" the dialogue history (what has been discussed previously) and the current question (what is asked), in addition to grounding information in the image, to accurately generate the correct response. In this paper, we show that co-attention models which explicitly encode dialoh history outperform models that don't, achieving state-of-the-art performance (72 % NDCG on val set). However, we also expose shortcomings of the crowdsourcing dataset collection procedure, by showing that dialogue history is indeed only required for a small amount of the data, and that the current evaluation metric encourages generic replies. To that end, we propose a challenging subset (VisdialConv) of the VisdialVal set and the benchmark NDCG of 63%.

Knowledge Supports Visual Language Grounding: A Case Study on Colour Terms [Website][PDF]
Simeon Schüz and Sina Zarrieß 20:00–21:00

In human cognition, world knowledge supports the perception of object colours: knowing that trees are typically green helps to perceive their colour in certain contexts. We go beyond previous studies on colour terms using isolated colour swatches and study visual grounding of colour terms in realistic objects. Our models integrate processing of visual information and object-specific knowledge via hard-coded (late) or learned (early) fusion. We find that both models consistently outperform a bottom-up baseline that predicts colour terms solely from visual inputs, but show interesting differences when predicting atypical colours of so-called colour diagnostic objects. Our models also achieve promising results when tested on new object categories not seen during training.

Learning Web-based Procedures by Reasoning over Explanations and Demonstrations in Context [Website][PDF]
Shashank Srivastava, Oleksandr Polozov, Nebojsa Jojic, and Christopher Meek 20:00–21:00

We explore learning web-based tasks from a human teacher through natural language explanations and a single demonstration. Our approach investigates a new direction for semantic parsing that models explaining a demonstration in a context, rather than mapping explanations to demonstrations. By leveraging the idea of inverse semantics from program synthesis to reason backwards from observed demonstrations, we ensure that all considered interpretations are consistent with executable actions in any context, thus simplifying the problem of search over logical forms. We present a dataset of explanations paired with demonstrations for web-based tasks. Our methods show better task completion rates than a supervised semantic parsing baseline (40% relative improvement on average), and are competitive with simple exploration-and-demonstration based methods, while requiring no exploration of the environment. In learning to align explanations with demonstrations, basic properties of natural language syntax emerge as learned behavior. This is an interesting example of pragmatic language acquisition without any linguistic annotation.

Mapping Natural Language Instructions to Mobile UI Action Sequences [Website][PDF]
Yang Li, Jiacong He, Xin Zhou, Yuan Zhang, and Jason Baldridge 20:00–21:00

We present a new problem: grounding natural language instructions to mobile user interface actions, and create three new datasets for it. For full task evaluation, we create PixelHelp, a corpus that pairs English instructions with actions performed by people on a mobile UI emulator. To scale training, we decouple the language and action data by (a) annotating action phrase spans in How-To instructions and (b) synthesizing grounded descriptions of actions for mobile user interfaces. We use a Transformer to extract action phrase tuples from long-range natural language instructions. A grounding Transformer then contextually represents UI objects using both their content and screen position and

connects them to object descriptions. Given a starting screen and instruction, our model achieves 70.59% accuracy on predicting complete ground-truth action sequences in PixelHelp.

Refer360°: A Referring Expression Recognition Dataset in 360° Images [Website][PDF]
Volkan Cirik, Taylor Berg-Kirkpatrick, and Louis-Philippe Morency 20:00–21:00

We propose a novel large-scale referring expression recognition dataset, Refer360°, consisting of 17,137 instruction sequences and ground-truth actions for completing these instructions in 360° scenes. Refer360° differs from existing related datasets in three ways. First, we propose a more realistic scenario where instructors and the followers have partial, yet dynamic, views of the scene – followers continuously modify their field-of-view (FoV) while interpreting instructions that specify a final target location. Second, instructions to find the target location consist of multiple steps for followers who will start at random FoVs. As a result, intermediate instructions are strongly grounded in object references, and followers must identify intermediate FoVs to find the final target location correctly. Third, the target locations are neither restricted to predefined objects nor chosen by annotators; instead, they are distributed randomly across scenes. This "point anywhere" approach leads to more linguistically complex instructions, as shown in our analyses. Our examination of the dataset shows that Refer360° manifests linguistically rich phenomena in a language grounding task that poses novel challenges for computational modeling of language, vision, and navigation.

TVQA+: Spatio-Temporal Grounding for Video Question Answering [Website][PDF]
Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal 20:00–21:00

We present the task of Spatio-Temporal Video Question Answering, which requires intelligent systems to simultaneously retrieve relevant moments and detect referenced visual concepts (people and objects) to answer natural language questions about videos. We first augment the TVQA dataset with 310.8K bounding boxes, linking depicted objects to visual concepts in questions and answers. We name this augmented version as TVQA+. We then propose Spatio-Temporal Answerer with Grounded Evidence (STAGE), a unified framework that grounds evidence in both spatial and temporal domains to answer questions about videos. Comprehensive experiments and analyses demonstrate the effectiveness of our framework and how the rich annotations in our TVQA+ dataset can contribute to the question answering task. Moreover, by performing this joint task, our model is able to produce insightful and interpretable spatio-temporal attention visualizations.

Unsupervised Multimodal Neural Machine Translation with Pseudo Visual Pivoting [Website][PDF]
Po-Yao Huang, Junjie Hu, Xiaojun Chang, and Alexander Hauptmann 20:00–21:00

Unsupervised machine translation (MT) has recently achieved impressive results with monolingual corpora only. However, it is still challenging to associate source-target sentences in the latent space. As people speak different languages biologically share similar visual systems, the potential of achieving better alignment through visual content is promising yet under-explored in unsupervised multimodal MT (MMT). In this paper, we investigate how to utilize visual content for disambiguation and promoting latent space alignment in unsupervised MMT. Our model employs multimodal back-translation and features pseudo visual pivoting in which we learn a shared multilingual visual-semantic embedding space and incorporate visually-pivoted captioning as additional weak supervision. The experimental results on the widely used Multi30K dataset show that the proposed model significantly improves over the state-of-the-art methods and generalizes well when images are not available at the testing time.

Words Aren't Enough, Their Order Matters: On the Robustness of Grounding Visual Referring Expressions [Website][PDF]
Arjun Akula, Spandana Gella, Yaser Al-Onaizan, Song-Chun Zhu, and Siva Reddy 20:00–21:00

Visual referring expression recognition is a challenging task that requires natural language understanding in the context of an image. We critically examine RefCOCOg, a standard benchmark for this task, using a human study and show that 83.7% of test instances do not require reasoning on linguistic structure, i.e., words are enough to identify the target object, the word order doesn't matter. To measure the true progress of existing models, we split the test set into two sets, one which requires reasoning on linguistic structure and the other which doesn't. Additionally, we create an out-of-distribution dataset Ref-Adv by asking crowdworkers to perturb in-domain examples such that the target object changes. Using these datasets, we empirically show that existing methods fail to exploit linguistic structure and are 12% to 23% lower in performance than the established progress for this task. We also propose two methods, one based on contrastive learning and the other based on multi-task learning, to increase the robustness of ViLBERT, the current state-of-the-art model for this task. Our datasets are publicly available at https://github.com/aws/aws-refcocog-adv.

Session 15A: Machine Translation-17

Addressing Posterior Collapse with Mutual Information for Improved Variational Neural Machine Translation [Website][PDF]
Arya D. McCarthy, Xian Li, Jiatao Gu, and Ning Dong 20:00–21:00

This paper proposes a simple and effective approach to address the problem of posterior collapse in conditional variational autoencoders (CVAEs). It thus improves performance of machine translation models that use noisy or monolingual data, as well as in conventional settings. Extending Transformer and conditional VAEs, our proposed latent variable model measurably prevents posterior collapse by (1) using a modified evidence lower bound (ELBO) objective which promotes mutual information between the latent variable and the target, and (2) guiding the latent variable with an auxiliary bag-of-words prediction task. As a result, the proposed model yields improved translation quality compared to existing variational NMT models on WMT Ro↔En and De↔En. With latent variables being effectively utilized, our model demonstrates improved robustness over non-latent Transformer in handling uncertainty: exploiting noisy source-side monolingual data (up to +3.2 BLEU), and training with weakly aligned web-mined parallel data (up to +4.7 BLEU).

Evaluating Robustness to Input Perturbations for Neural Machine Translation [Website][PDF]
Xing Niu, Prashant Mathur, Georgiana Dinu, and Yaser Al-Onaizan 20:00–21:00

Neural Machine Translation (NMT) models are sensitive to small perturbations in the input. Robustness to such perturbations is typically measured using translation quality metrics such as BLEU on the noisy input. This paper proposes additional metrics which measure the relative degradation and changes in translation when small perturbations are added to the input. We focus on a class of models employing subword regularization to address robustness and perform extensive evaluations of these models using the robustness measures proposed. Results show that our proposed metrics reveal a clear trend of improved robustness to perturbations when subword regularization methods are used.

Hard-Coded Gaussian Attention for Neural Machine Translation [Website][PDF]
Weiqiu You, Simeng Sun, and Mohit Iyyer 20:00–21:00

Recent work has questioned the importance of the Transformer's multi-headed attention for achieving high translation quality. We push further in this direction by developing a "hard-coded" attention variant without any learned parameters. Surprisingly, replacing all learned self-attention heads in the encoder and decoder with fixed, input-agnostic Gaussian distributions minimally impacts BLEU scores across four different language pairs. However, additionally, hard-coding cross attention (which connects the decoder to the encoder) significantly lowers BLEU, suggesting that it is more important than self-attention. Much of this BLEU drop can be recovered by adding just a single learned cross attention head to an otherwise hard-coded Transformer. Taken as a whole, our results offer insight into which components of the Transformer are actually important, which we hope will guide future work into the development of simpler and more efficient attention-based models.

Learning a Multi-Domain Curriculum for Neural Machine Translation [Website][PDF]
Wei Wang, Ye Tian, Jiquan Ngiam, Yinfei Yang, Isaac Caswell, and Zarana Parekh 20:00–21:00

Most data selection research in machine translation focuses on improving a single domain. We perform data selection for multiple domains at once. This is achieved by carefully introducing instance-level domain-relevance features and automatically constructing a training curriculum to gradually concentrate on multi-domain relevant and noise-reduced data batches. Both the choice of features and the use of curriculum are crucial for balancing and improving all domains, including out-of-domain. In large-scale experiments, the multi-domain curriculum simultaneously reaches or outperforms the individual performance and brings solid gains over no-curriculum training.

Tagged Back-translation Revisited: Why Does It Really Work? [Website][PDF]
Benjamin Marie, Raphael Rubino, and Atsushi Fujita 20:00–21:00

In this paper, we show that neural machine translation (NMT) systems trained on large back-translated data overfit some of the characteristics of machine-translated texts. Such NMT systems better translate human-produced translations, i.e., translationese, but may largely worsen the translation quality of original texts. Our analysis reveals that adding a simple tag to back-translations prevents this quality degradation and improves on average the overall translation quality by helping the NMT system to distinguish back-translated data from original parallel data during training. We also show that, in contrast to high-resource configurations, NMT systems trained in low-resource settings are much less vulnerable to overfit back-translations. We conclude that the back-translations in the training data should always be tagged especially when the origin of the text to be translated is unknown.

Session 15A Semantics: Sentence Level-10

Active Learning for Coreference Resolution using Discrete Annotation [Website][PDF]
Belinda Z. Li, Gabriel Stanovsky, and Luke Zettlemoyer 20:00–21:00

We improve upon pairwise annotation for active learning in coreference resolution, by asking annotators to identify mention antecedents if a presented mention pair is deemed not coreferent. This simple modification, when combined with a novel mention clustering algorithm for selecting which examples to label, is much more efficient in terms of the performance obtained per annotation budget. In experiments with existing benchmark coreference datasets, we show that the signal from this additional question leads to significant performance gains per human-annotation hour. Future work can use our annotation protocol to effectively develop coreference models for new domains. Our code is publicly available.

[TACL] Decoding Brain Activity Associated with Literal and Metaphoric Sentence Comprehension using Distributional Semantic Models [Website][PDF]
Vesna G. Djokic, Jean Maillard, Luana Bulat, and Ekaterina Shutova 20:00–21:00

Recent years have seen a growing interest within the natural language processing (NLP) community in evaluating the ability of semantic models to capture human meaning representation in the brain. Existing research has mainly focused on applying semantic models to decode brain activity patterns associated with the meaning of individual words, and, more recently, this approach has been extended to sentences and larger text fragments. Our work is the first to investigate metaphor processing in the brain in this context. We evaluate a range of semantic models (word embeddings, compositional, and visual models) in their ability to decode brain activity associated with reading of both literal and metaphoric sentences. Our results suggest that compositional models and word embeddings are able to capture differences in the processing of literal and metaphoric sentences, providing support for the idea that the literal meaning is not fully accessible during familiar metaphor comprehension.

Emerging Cross-lingual Structure in Pretrained Language Models [Website][PDF]
Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov 20:00–21:00

We study the problem of multilingual masked language modeling, i.e. the training of a single model on concatenated text from multiple languages, and present a detailed study of several factors that influence why these models are so effective for cross-lingual transfer. We show, contrary to what was previously hypothesized, that transfer is possible even when there is no shared vocabulary across the monolingual corpora and also when the text comes from very different domains. The only requirement is that there are some shared parameters in the top layers of the multilingual encoder. To better understand this result, we also show that representations from monolingual BERT models in different languages can be aligned post-hoc quite effectively, strongly suggesting that, much like for non-contextual word embeddings, there are universal latent symmetries in the learned embedding spaces. For multilingual masked language modeling, these symmetries are automatically discovered and aligned during the joint training process.

Estimating Mutual Information Between Dense Word Embeddings [Website][PDF]
Vitalii Zhelezniak, Aleksandar Savkov, and Nils Hammerla 20:00–21:00

Word embedding-based similarity measures are currently among the top-performing methods on unsupervised semantic textual similarity (STS) tasks. Recent work has increasingly adopted a statistical view on these embeddings, with some of the top approaches being essentially various correlations (which include the famous cosine similarity). Another excellent candidate for a similarity measure is mutual information (MI), which can capture arbitrary dependencies between the variables and has a simple and intuitive expression. Unfortunately, its use in the context of dense word embeddings has so far been avoided due to difficulties with estimating MI for continuous data. In this work we go through a vast literature on estimating MI in such cases and single out the most promising methods, yielding a simple and elegant similarity measure for word embeddings. We show that mutual information is a viable alternative to correlations, gives an excellent signal that correlates well with human judgements of similarity and rivals existing state-of-the-art unsupervised methods.

Exploring Unexplored Generalization Challenges for Cross-Database Semantic Parsing [Website][PDF]
Alane Suhr, Ming-Wei Chang, Peter Shaw, and Kenton Lee 20:00–21:00

We study the task of cross-database semantic parsing (XSP), where a system that maps natural language utterances to executable SQL queries is evaluated on databases unseen during training. Recently, several datasets, including Spider, were proposed to support development of XSP systems. We propose a challenging evaluation setup for cross-database semantic parsing, focusing on variation across database schemas and in-domain language use. We re-purpose eight semantic parsing datasets that have been well-studied in the setting where in-domain training data is available, and instead use them as additional evaluation data for XSP systems instead. We build a system that performs well on Spider, and find that it struggles to generalize to our re-purposed set. Our setup uncovers several generalization challenges for cross-database semantic parsing, demonstrating the need to use and develop diverse training and evaluation datasets.

Good-Enough Compositional Data Augmentation [Website][PDF]
Jacob Andreas 20:00–21:00

We propose a simple data augmentation protocol aimed at providing a compositional inductive bias in conditional and unconditional sequence models. Under this protocol, synthetic training examples are constructed by taking real training examples and replacing (possibly discontinuous) fragments with other fragments that appear in at least one similar environment. The protocol is model-agnostic and useful for a variety of tasks. Applied to neural sequence-to-

sequence models, it reduces error rate by as much as 87% on diagnostic tasks from the SCAN dataset and 16% on a semantic parsing task. Applied to n-gram language models, it reduces perplexity by roughly 1% on small corpora in several languages.

Incorporating External Knowledge through Pre-training for Natural Language to Code Generation [Website][PDF]
Frank F. Xu, Zhengbao Jiang, Pengcheng Yin, Bogdan Vasilescu, and Graham Neubig 20:00–21:00

Open-domain code generation aims to generate code in a general-purpose programming language (such as Python) from natural language (NL) intents. Motivated by the intuition that developers usually retrieve resources on the web when writing code, we explore the effectiveness of incorporating two varieties of external knowledge into NL-to-code generation: automatically mined NL-code pairs from the online programming QA forum StackOverflow and programming language API documentation. Our evaluations show that combining the two sources with data augmentation and retrieval-based data re-sampling improves the current state-of-the-art by up to 2.2% absolute BLEU score on the code generation testbed CoNaLa. The code and resources are available at https://github.com/neulab/external-knowledge-codegen.

Predicting the Focus of Negation: Model and Error Analysis [Website][PDF]
Md Mosharaf Hossain, Kathleen Hamilton, Alexis Palmer, and Eduardo Blanco 20:00–21:00

The focus of a negation is the set of tokens intended to be negated, and a key component for revealing affirmative alternatives to negated utterances. In this paper, we experiment with neural networks to predict the focus of negation. Our main novelty is leveraging a scope detector to introduce the scope of negation as an additional input to the network. Experimental results show that doing so obtains the best results to date. Additionally, we perform a detailed error analysis providing insights into the main error categories, and analyze errors depending on whether the model takes into account scope and context information.

TaBERT: Pretraining for Joint Understanding of Textual and Tabular Data [Website][PDF]
Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel 20:00–21:00

Recent years have witnessed the burgeoning of pretrained language models (LMs) for text-based natural language (NL) understanding tasks. Such models are typically trained on free-form NL text, hence may not be suitable for tasks like semantic parsing over structured data, which require reasoning over both free-form NL questions and structured tabular data (e.g., database tables). In this paper we present TaBERT, a pretrained LM that jointly learns representations for NL sentences and (semi-)structured tables. TaBERT is trained on a large corpus of 26 million tables and their English contexts. In experiments, neural semantic parsers using TaBERT as feature representation layers achieve new best results on the challenging weakly-supervised semantic parsing benchmark WikiTableQuestions, while performing competitively on the text-to-SQL dataset Spider.

Unsupervised Cross-lingual Representation Learning at Scale [Website][PDF]
Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov 20:00–21:00

This paper shows that pretraining multilingual language models at scale leads to significant performance gains for a wide range of cross-lingual transfer tasks. We train a Transformer-based masked language model on one hundred languages, using more than two terabytes of filtered CommonCrawl data. Our model, dubbed XLM-R, significantly outperforms multilingual BERT (mBERT) on a variety of cross-lingual benchmarks, including +14.6% average accuracy on XNLI, +13% average F1 score on MLQA, and +2.4% F1 score on NER. XLM-R performs particularly well on low-resource languages, improving 15.7% in XNLI accuracy for Swahili and 11.4% for Urdu over previous XLM models. We also present a detailed empirical analysis of the key factors that are required to achieve these gains, including the trade-offs between (1) positive transfer and capacity dilution and (2) the performance of high and low resource languages at scale. Finally, we show, for the first time, the possibility of multilingual modeling without sacrificing per-language performance; XLM-R is very competitive with strong monolingual models on the GLUE and XNLI benchmarks. We will make our code and models publicly available.

Session 15A Semantics: Textual Inference and Other Areas of Semantics-7

Are Natural Language Inference Models IMPPRESsive? Learning IMPlicature and PRESupposition [Website][PDF]
Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams 20:00–21:00

Natural language inference (NLI) is an increasingly important task for natural language understanding, which requires one to infer whether a sentence entails another. However, the ability of NLI models to make pragmatic inferences remains understudied. We create an IMPlicature and PRESupposition diagnostic dataset (IMPPRES), consisting of 32K semi-automatically generated sentence pairs illustrating well-studied pragmatic inference types. We use IMPPRES to evaluate whether BERT, InferSent, and BOW NLI models trained on MultiNLI (Williams et al., 2018) learn to make pragmatic inferences. Although MultiNLI appears to contain very few pairs illustrating these inference types, we find that BERT learns to draw pragmatic inferences. It reliably treats scalar implicatures triggered by "some" as entailments. For some presupposition triggers like "only", BERT reliably recognizes the presupposition as an entailment, even when the trigger is embedded under an entailment canceling operator like negation. BOW and InferSent show weaker evidence of pragmatic reasoning. We conclude that NLI training encourages models to learn some, but not all, pragmatic inferences.

End-to-End Bias Mitigation by Modelling Biases in Corpora [Website][PDF]
Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson 20:00–21:00

Several recent studies have shown that strong natural language understanding (NLU) models are prone to relying on unwanted dataset biases without learning the underlying task, resulting in models that fail to generalize to out-of-domain datasets and are likely to perform poorly in real-world scenarios. We propose two learning strategies to train neural models, which are more robust to such biases and transfer better to out-of-domain datasets. The biases are specified in terms of one or more bias-only models, which learn to leverage the dataset biases. During training, the bias-only models' predictions are used to adjust the loss of the base model to reduce its reliance on biases by down-weighting the biased examples and focusing the training on the hard examples. We experiment on large-scale natural language inference and fact verification benchmarks, evaluating on out-of-domain datasets that are specifically designed to assess the robustness of models against known biases in the training data. Results show that our debiasing methods greatly improve robustness in all settings and better transfer to other textual entailment datasets. Our code and data are publicly available in `https://github.com/rabeehk/robust-nli`.

Mind the Trade-off: Debiasing NLU Models without Degrading the In-distribution Performance [Website][PDF]
Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych 20:00–21:00

Models for natural language understanding (NLU) tasks often rely on the idiosyncratic biases of the dataset, which make them brittle against test cases outside the training distribution. Recently, several proposed debiasing methods are shown to be very effective in improving out-of-distribution performance. However, their improvements come at the expense of performance drop when models are evaluated on the in-distribution data, which contain examples with higher diversity. This seemingly inevitable trade-off may not tell us much about the changes in the reasoning and understanding capabilities of the resulting models on broader types of examples beyond the small subset represented in the out-of-distribution data. In this paper, we address this trade-off by introducing a novel debiasing method, called confidence regularization, which discourage models from exploiting biases while enabling them to receive enough incentive to learn from all the training examples. We evaluate our method on three NLU tasks and show that, in contrast to its predecessors, it improves the performance on out-of-distribution datasets (e.g., 7pp gain on HANS dataset) while maintaining the original in-distribution accuracy.

NILE : Natural Language Inference with Faithful Natural Language Explanations [Website][PDF]
Sawan Kumar and Partha Talukdar 20:00–21:00

The recent growth in the popularity and success of deep learning models on NLP classification tasks has accompanied the need for generating some form of natural language explanation of the predicted labels. Such generated natural language (NL) explanations are expected to be faithful, i.e., they should correlate well with the model's internal decision making. In this work, we focus on the task of natural language inference (NLI) and address the following question: can we build NLI systems which produce labels with high accuracy, while also generating faithful explanations of its decisions? We propose Natural-language Inference over Label-specific Explanations (NILE), a novel NLI method which utilizes auto-generated label-specific NL explanations to produce labels along with its faithful explanation. We demonstrate NILE's effectiveness over previously reported methods through automated and human evaluation of the produced labels and explanations. Our evaluation of NILE also supports the claim that accurate systems capable of providing testable explanations of their decisions can be designed. We discuss the faithfulness of NILE's explanations in terms of sensitivity of the decisions to the corresponding explanations. We argue that explicit evaluation of faithfulness, in addition to label and explanation accuracy, is an important step in evaluating model's explanations. Further, we demonstrate that task-specific probes are necessary to establish such sensitivity.

QuASE: Question-Answer Driven Sentence Encoding [Website][PDF]
Hangfeng He, Qiang Ning, and Dan Roth 20:00–21:00

Question-answering (QA) data often encodes essential information in many facets. This paper studies a natural question: Can we get supervision from QA data for other tasks (typically, non-QA ones)? For example, *can we use QAMR (Michael et al., 2017) to improve named entity recognition?* We suggest that simply further pre-training BERT is often not the best option, and propose the *question-answer driven sentence encoding (QuASE)* framework. QuASE learns representations from QA data, using BERT or other state-of-the-art contextual language models. In particular, we

observe the need to distinguish between two types of sentence encodings, depending on whether the target task is a single- or multi-sentence input; in both cases, the resulting encoding is shown to be an easy-to-use plugin for many downstream tasks. This work may point out an alternative way to supervise NLP tasks.

Towards Robustifying NLI Models Against Lexical Dataset Biases [Website] [PDF]
Xiang Zhou and Mohit Bansal 20:00–21:00

While deep learning models are making fast progress on the task of Natural Language Inference, recent studies have also shown that these models achieve high accuracy by exploiting several dataset biases, and without deep understanding of the language semantics. Using contradiction-word bias and word-overlapping bias as our two bias examples, this paper explores both data-level and model-level debiasing methods to robustify models against lexical dataset biases. First, we debias the dataset through data augmentation and enhancement, but show that the model bias cannot be fully removed via this method. Next, we also compare two ways of directly debiasing the model without knowing what the dataset biases are in advance. The first approach aims to remove the label bias at the embedding level. The second approach employs a bag-of-words sub-model to capture the features that are likely to exploit the bias and prevents the original model from learning these biased features by forcing orthogonality between these two sub-models. We performed evaluations on new balanced datasets extracted from the original MNLI dataset as well as the NLI stress tests, and show that the orthogonality approach is better at debiasing the model while maintaining competitive overall accuracy.

Uncertain Natural Language Inference [Website] [PDF]
Tongfei Chen, Zhengping Jiang, Adam Poliak, Keisuke Sakaguchi, and Benjamin Van Durme 20:00–21:00

We introduce Uncertain Natural Language Inference (UNLI), a refinement of Natural Language Inference (NLI) that shifts away from categorical labels, targeting instead the direct prediction of subjective probability assessments. We demonstrate the feasibility of collecting annotations for UNLI by relabeling a portion of the SNLI dataset under a probabilistic scale, where items even with the same categorical label differ in how likely people judge them to be true given a premise. We describe a direct scalar regression modeling approach, and find that existing categorically-labeled NLI data can be used in pre-training. Our best models correlate well with humans, demonstrating models are capable of more subtle inferences than the categorical bin assignment employed in current NLI tasks.

Session 15A: Student Research Workshop

Checkpoint Reranking: An Approach to Select Better Hypothesis for Neural Machine Translation Systems [Website][PDF]
Vinay Pandramish and Dipti Misra Sharma 20:00–21:00

In this paper, we propose a method of re-ranking the outputs of Neural Machine Translation (NMT) systems. After the decoding process, we select a few last iteration outputs in the training process as the N-best list. After training a Neural Machine Translation (NMT) baseline system, it has been observed that these iteration outputs have an oracle score higher than baseline up to 1.01 BLEU points compared to the last iteration of the trained system.We come up with a ranking mechanism by solely focusing on the decoder's ability to generate distinct tokens and without the usage of any language model or data. With this method, we achieved a translation improvement up to +0.16 BLEU points over baseline.We also evaluate our approach by applying the coverage penalty to the training process.In cases of moderate coverage penalty, the oracle scores are higher than the final iteration up to +0.99 BLEU points, and our algorithm gives an improvement up to +0.17 BLEU points.With excessive penalty, there is a decrease in translation quality compared to the baseline system. Still, an increase in oracle scores up to +1.30 is observed with the re-ranking algorithm giving an improvement up to +0.15 BLEU points is found in case of excessive penalty.The proposed re-ranking method is a generic one and can be extended to other language pairs as well.

Story-level Text Style Transfer: A Proposal [Website][PDF]
Yusu Qian 20:00–21:00

Text style transfer aims to change the style of the input text to the target style while preserving the content to some extent. Previous works on this task are on the sentence level. We aim to work on story-level text style transfer to generate stories that preserve the plot of the input story while exhibiting a strong target style. The challenge in this task compared to previous work is that the structure of the input story, consisting of named entities and their relations with each other, needs to be preserved, and that the generated story needs to be consistent after adding flavors. We plan to explore three methods including the BERT-based method, the Story Realization method, and the Graph-based method.

Non-Topical Coherence in Social Talk: A Call for Dialogue Model Enrichment [Website][PDF]
Alex Luu and Sophia A. Malamud 20:00–21:00

Current models of dialogue mainly focus on utterances within a topically coherent discourse segment, rather than new-topic utterances (NTUs), which begin a new topic not correlating with the content of prior discourse. As a result, these models may sufficiently account for discourse context of task-oriented but not social conversations. We conduct a pilot annotation study of NTUs as a first step towards a model capable of rationalizing conversational coherence in social talk. We start with the naturally occurring social dialogues in the Disco-SPICE corpus, annotated with discourse relations in the Penn Discourse Treebank and Cognitive approach to Coherence Relations frameworks. We first annotate content-based coherence relations that are not available in Disco-SPICE, and then heuristically identify NTUs, which lack a coherence relation to prior discourse. Based on the interaction between NTUs and their discourse context, we construct a classification for NTUs that actually convey certain non-topical coherence in social talk. This classification introduces new sequence-based social intents that traditional taxonomies of speech acts do not capture. The new findings advocates the development of a Bayesian game-theoretic model for social talk.

Compositional Generalization by Factorizing Alignment and Translation [Website][PDF]
Jacob Russin, Jason Jo, Randall O'Reilly, and Yoshua Bengio 20:00–21:00

Standard methods in deep learning for natural language processing fail to capture the compositional structure of human language that allows for systematic generalization outside of the training distribution. However, human learners readily generalize in this way, e.g. by applying known grammatical rules to novel words. Inspired by work in cognitive science suggesting a functional distinction between systems for syntactic and semantic processing, we implement a modification to an existing approach in neural machine translation, imposing an analogous separation between alignment and translation. The resulting architecture substantially outperforms standard recurrent networks on the SCAN dataset, a compositional generalization task, without any additional supervision. Our work suggests that learning to align and to translate in separate modules may be a useful heuristic for capturing compositional structure.

Demo Session 5B

Time: *20:45–21:30*

NSTM: Real-Time Query-Driven News Overview Composition at Bloomberg [Website][PDF]
Joshua Bambrick, Minjie Xu, Andy Almonte, Igor Malioutov, Guim Perarnau, Vittorio Selo, and Iat Chong Chan

Millions of news articles from hundreds of thousands of sources around the globe appear in news aggregators every day. Consuming such a volume of news presents an almost insurmountable challenge. For example, a reader searching on Bloomberg's system for news about the U.K. would find 10,000 articles on a typical day. Apple Inc., the world's most journalistically covered company, garners around 1,800 news articles a day. We realized that a new kind of summarization engine was needed, one that would condense large volumes of news into short, easy to absorb points. The system would filter out noise and duplicates to identify and summarize key news about companies, countries or markets. When given a user query, Bloomberg's solution, Key News Themes (or NSTM), leverages state-of-the-art semantic clustering techniques and novel summarization methods to produce comprehensive, yet concise, digests to dramatically simplify the news consumption process. NSTM is available to hundreds of thousands of readers around the world and serves thousands of requests daily with sub-second latency. At ACL 2020, we will present a demo of NSTM.

LEAN-LIFE: A Label-Efficient Annotation Framework Towards Learning from Explanation [Website][PDF]
Dong-Ho Lee, Rahul Khanna, Bill Yuchen Lin, Seyeon Lee, Qinyuan Ye, Elizabeth Boschee, Leonardo Neves, and Xiang Ren

Successfully training a deep neural network demands a huge corpus of labeled data. However, each label only provides limited information to learn from, and collecting the requisite number of labels involves massive human effort. In this work, we introduce LEAN-LIFE, a web-based, Label-Efficient AnnotatioN framework for sequence labeling and classification tasks, with an easy-to-use UI that not only allows an annotator to provide the needed labels for a task but also enables LearnIng From Explanations for each labeling decision. Such explanations enable us to generate useful additional labeled data from unlabeled instances, bolstering the pool of available training data. On three popular NLP tasks (named entity recognition, relation extraction, sentiment analysis), we find that using this enhanced supervision allows our models to surpass competitive baseline F1 scores by more than 5-10 percentage points, while using 2X times fewer labeled instances. Our framework is the first to utilize this enhanced supervision technique and does so for three important tasks – thus providing improved annotation recommendations to users and an ability to build datasets of (data, label, explanation) triples instead of the regular (data, label) pair.

Session 15B Overview – Wednesday, July 8, 2020 21:00–22:00

Track A *Generation-15* Abstracts	ESPRIT: Explaining Solutions to Physical Reasoning Tasks *Rajani, Zhang, Tan, Zheng, Weiss, Vyas, Gupta, Xiong, Socher, and Radev* [Website][PDF]	Logical Natural Language Generation from Open-Domain Tables *Chen, Chen, Su, Chen, and Wang* [Website][PDF]	R^3: Reverse, Retrieve, and Rank for Sarcasm Generation with Commonsense Knowledge *Chakrabarty, Ghosh, Muresan, and Peng* [Website][PDF]	Structural Information Preserving for Graph-to-Text Generation *Song, Wang, Su, Zhang, Xu, Ge, and Yu* [Website][PDF]	Toward Better Storylines with Sentence-Level Language Models *Ippolito, Grangier, Eck, and Callison-Burch* [Website][PDF]
Track B *Information Extraction-12* Abstracts	Exploiting the Syntax-Model Consistency for Neural Relation Extraction *Pouran Ben Veyseh, Dernoncourt, Dou, and Nguyen* [Website][PDF]	From English to Code-Switching: Transfer Learning with Strong Morphological Clues *Aguilar and Solorio* [Website][PDF]	Hierarchical Entity Typing via Multi-level Learning to Rank *Chen, Chen, and Van Durme* [Website][PDF]	Machine Reading of Historical Events *Honovich, Torroba Hennigen, Abend, and Cohen* [Website][PDF]	Multi-Domain Named Entity Recognition with Genre-Aware and Agnostic Inference *Wang, Kulkarni, and Preotiuc-Pietro* [Website][PDF]
	Rationalizing Medical Relation Prediction from Corpus-level Statistics *Wang, Lee, Lin, and Sun* [Website][PDF]	Representation Learning for Information Extraction from Form-like Documents *Majumder, Potti, Tata, Wendt, Zhao, and Najork* [Website][PDF]	SciREX: A Challenge Dataset for Document-Level Information Extraction *Jain, Zuylen, Hajishirzi, and Beltagy* [Website][PDF]	Soft Gazetteers for Low-Resource Named Entity Recognition *Rijhwani, Zhou, Neubig, and Carbonell* [Website][PDF]	Sources of Transfer in Multilingual Named Entity Recognition *Mueller, Andrews, and Dredze* [Website][PDF]
	TXtract: Taxonomy-Aware Knowledge Extraction for Thousands of Product Categories *Karamanolakis, Ma, and Dong* [Website][PDF]	TriggerNER: Learning with Entity Triggers as Explanations for Named Entity Recognition *Lin, Lee, Shen, Moreno, Huang, Shiralkar, and Ren* [Website][PDF]			
Track C *Machine Translation-18* Abstracts	Balancing Training for Multilingual Neural Machine Translation *Wang, Tsvetkov, and Neubig* [Website][PDF]	HAT: Hardware-Aware Transformers for Efficient Natural Language Processing *Wang, Wu, Liu, Cai, Zhu, Gan, and Han* [Website][PDF]	Parallel Corpus Filtering via Pre-trained Language Models *Zhang, Nagesh, and Knight* [Website][PDF]	Regularized Context Gates on Transformer for Machine Translation *Li, Liu, Wang, Huang, and Meng* [Website][PDF]	Translationese as a Language in "Multilingual" NMT *Riley, Caswell, Freitag, and Grangier* [Website][PDF]
Track D *NLP Applications-12* Abstracts	A Multi-Perspective Architecture for Semantic Code Search *Haldar, Wu, Xiong, and Hockenmaier* [Website][PDF]	Automated Topical Component Extraction Using Neural Network Attention Scores from Source-based Essay Scoring *Zhang and Litman* [Website][PDF]	Clinical Concept Linking with Contextualized Neural Representations *Schumacher, Mulyar, and Dredze* [Website][PDF]	DeSePtion: Dual Sequence Prediction and Adversarial Examples for Improved Fact-Checking *Hidey, Chakrabarty, Alhindi, Varia, Krstovski, Diab, and Muresan* [Website][PDF]	Let Me Choose: From Verbal Context to Font Selection *Shirani, Dernoncourt, Echevarria, Asente, Lipka, and Solorio* [Website][PDF]

	[TACL] Machine Learning Driven Language Assessment *Settles, Hagiwara, and LaFlair* [Website][PDF]	Multi-Label and Multilingual News Framing Analysis *Akyürek, Guo, Elanwar, Ishwar, Betke, and Wijaya* [Website][PDF]	Predicting Performance for Natural Language Processing Tasks *Xia, Anastasopoulos, Xu, Yang, and Neubig* [Website][PDF]	ScriptWriter: Narrative-Guided Script Generation *Zhu, Song, Dou, NIE, and Zhou* [Website][PDF]	Should All Cross-Lingual Embeddings Speak English? *Anastasopoulos and Neubig* [Website][PDF]
	Smart To-Do: Automatic Generation of To-Do Items from Emails *Mukherjee, Mukherjee, Hasegawa, Hassan Awadallah, and White* [Website][PDF]				
Track E *Phonology, Morphology and Word Segmentation-7* Abstracts	A Multitask Learning Approach for Diacritic Restoration *Alqahtani, Mishra, and Diab* [Website][PDF]	Frugal Paradigm Completion *Erdmann, Kenter, Becker, and Schallhart* [Website][PDF]	Improving Chinese Word Segmentation with Wordhood Memory Networks *Tian, Song, Xia, Zhang, and Wang* [Website][PDF]	Joint Chinese Word Segmentation and Part-of-speech Tagging via Two-way Attentions of Auto-analyzed Knowledge *Tian, Song, Ao, Xia, Quan, Zhang, and Wang* [Website][PDF]	Joint Diacritization, Lemmatization, Normalization, and Fine-Grained Morphological Tagging *Zalmout and Habash* [Website][PDF]
	Phonetic and Visual Priors for Decipherment of Informal Romanization *Ryskina, Gormley, and Berg-Kirkpatrick* [Website][PDF]	Unsupervised Morphological Paradigm Completion *Jin, Cai, Peng, Xia, McCarthy, and Kann* [Website][PDF]			
Track F *Sentence Level-11* Abstracts	Beyond Possession Existence: Duration and Co-Possession *Chinnappa, Murugan, and Blanco* [Website][PDF]	Controlled Crowdsourcing for High-Quality QA-SRL Annotation *Roit, Klein, Stepanov, Mamou, Michael, Stanovsky, Zettlemoyer, and Dagan* [Website][PDF]	Don't Stop Pretraining: Adapt Language Models to Domains and Tasks *Gururangan, Marasović, Swayamdipta, Lo, Beltagy, Downey, and Smith* [Website][PDF]	Structured Tuning for Semantic Role Labeling *Li, Jawale, Palmer, and Srikumar* [Website][PDF]	Universal Decompositional Semantic Parsing *Stengel-Eskin, White, Zhang, and Van Durme* [Website][PDF]
Track G *Student Research Workshop* Abstracts	#NotAWhore! A Computational Linguistic Perspective of Rape Culture and Victimization on Social Media *Suvarna and Bhalla* [Website][PDF]	Crossing the Line: Where do Demographic Variables Fit into Humor Detection? *Meaney* [Website][PDF]	Effectively Aligning and Filtering Parallel Corpora under Sparse Data Conditions *Steingrímsson, Loftsson, and Way* [Website][PDF]	Exploring the Role of Context to Distinguish Rhetorical and Information-Seeking Questions *Zhuang and Riloff* [Website][PDF]	
Track H *Tagging, Chunking and Parsing-6* Abstracts	[TACL] Deep Contextualized Self-training for Low Resource Dependency Parsing *Rotman and Reichart* [Website][PDF]	Extracting Headless MWEs from Dependency Parse Trees: Parsing, Tagging, and Joint Modeling Approaches *Shi and Lee* [Website][PDF]	Revisiting Higher-Order Dependency Parsers *Fonseca and Martins* [Website][PDF]	SeqVAT: Virtual Adversarial Training for Semi-Supervised Sequence Labeling *Chen, Ruan, Liu, and Lu* [Website][PDF]	Treebank Embedding Vectors for Out-of-domain Dependency Parsing *Wagner, Barry, and Foster* [Website][PDF]

Track I *Theme-6* Abstracts	Automated Evaluation of Writing — 50 Years and Counting *Beigman Klebanov and Madnani* [Website] [PDF]	Negated and Misprimed Probes for Pretrained Language Models: Birds Can Talk, But Cannot Fly *Kassner and Schütze* [Website] [PDF]	On Forgetting to Cite Older Papers: An Analysis of the ACL Anthology *Bollmann and Elliott* [Website] [PDF]	Returning the N to NLP: Towards Contextually Personalized Classification Models *Flek* [Website] [PDF]	The Unstoppable Rise of Computational Linguistics in Deep Learning *Henderson* [Website] [PDF]
	To Boldly Query What No One Has Annotated Before? The Frontiers of Corpus Querying *Gärtner and Jung* [Website] [PDF]	To Test Machine Comprehension, Start by Defining Comprehension *Dunietz, Burnham, Bharadwaj, Rambow, Chu-Carroll, and Ferrucci* [Website] [PDF]	Gender Gap in Natural Language Processing Research: Disparities in Authorship and Citations *Mohammad* [Website] [PDF]		

Session 15B Details

Session 15B: Generation-15

ESPRIT: Explaining Solutions to Physical Reasoning Tasks [Website][PDF]
Nazneen Fatema Rajani, Rui Zhang, Yi Chern Tan, Stephan Zheng, Jeremy Weiss, Aadit Vyas, Abhijit Gupta, Caiming Xiong, Richard Socher, and Dragomir Radev 21:00–22:00

Neural networks lack the ability to reason about qualitative physics and so cannot generalize to scenarios and tasks unseen during training. We propose ESPRIT, a framework for commonsense reasoning about qualitative physics in natural language that generates interpretable descriptions of physical events. We use a two-step approach of first identifying the pivotal physical events in an environment and then generating natural language descriptions of those events using a data-to-text approach. Our framework learns to generate explanations of how the physical simulation will causally evolve so that an agent or a human can easily reason about a solution using those interpretable descriptions. Human evaluations indicate that ESPRIT produces crucial fine-grained details and has high coverage of physical concepts compared to even human annotations. Dataset, code and documentation are available at `https://github.com/salesforce/esprit`.

Logical Natural Language Generation from Open-Domain Tables [Website][PDF]
Wenhu Chen, Jianshu Chen, Yu Su, Zhiyu Chen, and William Yang Wang 21:00–22:00

Neural natural language generation (NLG) models have recently shown remarkable progress in fluency and coherence. However, existing studies on neural NLG are primarily focused on surface-level realizations with limited emphasis on logical inference, an important aspect of human thinking and language. In this paper, we suggest a new NLG task where a model is tasked with generating natural language statements that can be *logically entailed* by the facts in an open-domain semi-structured table. To facilitate the study of the proposed logical NLG problem, we use the existing TabFact dataset~[**chen2019tabfact**] featured with a wide range of logical/symbolic inferences as our testbed, and propose new automatic metrics to evaluate the fidelity of generation models w.r.t. logical inference. The new task poses challenges to the existing monotonic generation frameworks due to the mismatch between sequence order and logical order. In our experiments, we comprehensively survey different generation architectures (LSTM, Transformer, Pre-Trained LM) trained with different algorithms (RL, Adversarial Training, Coarse-to-Fine) on the dataset and made following observations: 1) Pre-Trained LM can significantly boost both the fluency and logical fidelity metrics, 2) RL and Adversarial Training are trading fluency for fidelity, 3) Coarse-to-Fine generation can help partially alleviate the fidelity issue while maintaining high language fluency. The code and data are available at `https://github.com/wenhuchen/LogicNLG`.

R^3: Reverse, Retrieve, and Rank for Sarcasm Generation with Commonsense Knowledge [Website][PDF]
Tuhin Chakrabarty, Debanjan Ghosh, Smaranda Muresan, and Nanyun Peng 21:00–22:00

We propose an unsupervised approach for sarcasm generation based on a non-sarcastic input sentence. Our method employs a retrieve-and-edit framework to instantiate two major characteristics of sarcasm: reversal of valence and semantic incongruity with the context, which could include shared commonsense or world knowledge between the speaker and the listener. While prior works on sarcasm generation predominantly focus on context incongruity, we show that combining valence reversal and semantic incongruity based on the commonsense knowledge generates sarcasm of higher quality. Human evaluation shows that our system generates sarcasm better than humans 34% of the time, and better than a reinforced hybrid baseline 90% of the time.

Structural Information Preserving for Graph-to-Text Generation [Website][PDF]
Linfeng Song, Ante Wang, Jinsong Su, Yue Zhang, Kun Xu, Yubin Ge, and Dong Yu 21:00–22:00

The task of graph-to-text generation aims at producing sentences that preserve the meaning of input graphs. As a crucial defect, the current state-of-the-art models may mess up or even drop the core structural information of input graphs when generating outputs. We propose to tackle this problem by leveraging richer training signals that can guide our model for preserving input information. In particular, we introduce two types of autoencoding losses, each individually focusing on different aspects (a.k.a. views) of input graphs. The losses are then back-propagated to better calibrate our model via multi-task training. Experiments on two benchmarks for graph-to-text generation show the effectiveness of our approach over a state-of-the-art baseline.

Toward Better Storylines with Sentence-Level Language Models [Website][PDF]
Daphne Ippolito, David Grangier, Douglas Eck, and Chris Callison-Burch 21:00–22:00

We propose a sentence-level language model which selects the next sentence in a story from a finite set of fluent alternatives. Since it does not need to model fluency, the sentence-level language model can focus on longer range dependencies, which are crucial for multi-sentence coherence. Rather than dealing with individual words, our method treats the story so far as a list of pre-trained sentence embeddings and predicts an embedding for the next sentence, which is more efficient than predicting word embeddings. Notably this allows us to consider a large number of candidates for the next sentence during training. We demonstrate the effectiveness of our approach with state-of-the-art accuracy on the unsupervised Story Cloze task and with promising results on larger-scale next sentence prediction tasks.

Session 15B: Information Extraction-12

Exploiting the Syntax-Model Consistency for Neural Relation Extraction [Website][PDF]
Amir Pouran Ben Veyseh, Franck Dernoncourt, Dejing Dou, and Thien Huu Nguyen 21:00–22:00

This paper studies the task of Relation Extraction (RE) that aims to identify the semantic relations between two entity mentions in text. In the deep learning models for RE, it has been beneficial to incorporate the syntactic structures from the dependency trees of the input sentences. In such models, the dependency trees are often used to directly structure the network architectures or to obtain the dependency relations between the word pairs to inject the syntactic information into the models via multi-task learning. The major problem with these approaches is the lack of generalization beyond the syntactic structures in the training data or the failure to capture the syntactic importance of the words for RE. In order to overcome these issues, we propose a novel deep learning model for RE that uses the dependency trees to extract the syntax-based importance scores for the words, serving as a tree representation to introduce syntactic information into the models with greater generalization. In particular, we leverage Ordered-Neuron Long-Short Term Memory Networks (ON-LSTM) to infer the model-based importance scores for RE for every word in the sentences that are then regulated to be consistent with the syntax-based scores to enable syntactic information injection. We perform extensive experiments to demonstrate the effectiveness of the proposed method, leading to the state-of-the-art performance on three RE benchmark datasets.

From English to Code-Switching: Transfer Learning with Strong Morphological Clues [Website][PDF]
Gustavo Aguilar and Thamar Solorio 21:00–22:00

Linguistic Code-switching (CS) is still an understudied phenomenon in natural language processing. The NLP community has mostly focused on monolingual and multi-lingual scenarios, but little attention has been given to CS in particular. This is partly because of the lack of resources and annotated data, despite its increasing occurrence in social media platforms. In this paper, we aim at adapting monolingual models to code-switched text in various tasks. Specifically, we transfer English knowledge from a pre-trained ELMo model to different code-switched language pairs (i.e., Nepali-English, Spanish-English, and Hindi-English) using the task of language identification. Our method, CS-ELMo, is an extension of ELMo with a simple yet effective position-aware attention mechanism inside its character convolutions. We show the effectiveness of this transfer learning step by outperforming multilingual BERT and homologous CS-unaware ELMo models and establishing a new state of the art in CS tasks, such as NER and POS tagging. Our technique can be expanded to more English-paired code-switched languages, providing more resources to the CS community.

Hierarchical Entity Typing via Multi-level Learning to Rank [Website][PDF]
Tongfei Chen, Yunmo Chen, and Benjamin Van Durme 21:00–22:00

We propose a novel method for hierarchical entity classification that embraces ontological structure at both training and during prediction. At training, our novel multi-level learning-to-rank loss compares positive types against negative siblings according to the type tree. During prediction, we define a coarse-to-fine decoder that restricts viable candidates at each level of the ontology based on already predicted parent type(s). Our approach significantly outperform prior work on strict accuracy, demonstrating the effectiveness of our method.

Machine Reading of Historical Events [Website][PDF]
Or Honovich, Lucas Torroba Hennigen, Omri Abend, and Shay B. Cohen 21:00–22:00

Machine reading is an ambitious goal in NLP that subsumes a wide range of text understanding capabilities. Within this broad framework, we address the task of machine reading the time of historical events, compile datasets for the task, and develop a model for tackling it. Given a brief textual description of an event, we show that good performance can be achieved by extracting relevant sentences from Wikipedia, and applying a combination of task-specific and general-purpose feature embeddings for the classification. Furthermore, we establish a link between the historical event ordering task and the event focus time task from the information retrieval literature, showing they also provide a challenging test case for machine reading algorithms.

Multi-Domain Named Entity Recognition with Genre-Aware and Agnostic Inference [Website][PDF]
Jing Wang, Mayank Kulkarni, and Daniel Preotiuc-Pietro 21:00–22:00

Named entity recognition is a key component of many text processing pipelines and it is thus essential for this component to be robust to different types of input. However, domain transfer of NER models with data from multiple genres has not been widely studied. To this end, we conduct NER experiments in three predictive setups on data from: a) multiple domains; b) multiple domains where the genre label is unknown at inference time; c) domains not encountered in training. We introduce a new architecture tailored to this task by using shared and private domain parameters and multi-task learning. This consistently outperforms all other baseline and competitive methods on all three experimental setups, with differences ranging between +1.95 to +3.11 average F1 across multiple genres when compared to standard approaches. These results illustrate the challenges that need to be taken into account when building real-world NLP applications that are robust to various types of text and the methods that can help, at least partially, alleviate these issues.

Rationalizing Medical Relation Prediction from Corpus-level Statistics [Website][PDF]
Zhen Wang, Jennifer Lee, Simon Lin, and Huan Sun 21:00–22:00

Nowadays, the interpretability of machine learning models is becoming increasingly important, especially in the medical domain. Aiming to shed some light on how to rationalize medical relation prediction, we present a new interpretable framework inspired by existing theories on how human memory works, e.g., theories of recall and recog-

nition. Given the corpus-level statistics, i.e., a global co-occurrence graph of a clinical text corpus, to predict the relations between two entities, we first recall rich contexts associated with the target entities, and then recognize relational interactions between these contexts to form model rationales, which will contribute to the final prediction. We conduct experiments on a real-world public clinical dataset and show that our framework can not only achieve competitive predictive performance against a comprehensive list of neural baseline models, but also present rationales to justify its prediction. We further collaborate with medical experts deeply to verify the usefulness of our model rationales for clinical decision making.

Representation Learning for Information Extraction from Form-like Documents [Website][PDF]
Bodhisattwa Prasad Majumder, Navneet Potti, Sandeep Tata, James Bradley Wendt, Qi Zhao, and Marc Najork 21:00–22:00

We propose a novel approach using representation learning for tackling the problem of extracting structured information from form-like document images. We propose an extraction system that uses knowledge of the types of the target fields to generate extraction candidates and a neural network architecture that learns a dense representation of each candidate based on neighboring words in the document. These learned representations are not only useful in solving the extraction task for unseen document templates from two different domains but are also interpretable, as we show using loss cases.

SciREX: A Challenge Dataset for Document-Level Information Extraction [Website][PDF]
Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy 21:00–22:00

Extracting information from full documents is an important problem in many domains, but most previous work focus on identifying relationships within a sentence or a paragraph. It is challenging to create a large-scale information extraction (IE) dataset at the document level since it requires an understanding of the whole document to annotate entities and their document-level relationships that usually span beyond sentences or even sections. In this paper, we introduce SciREX, a document level IE dataset that encompasses multiple IE tasks, including salient entity identification and document level N-ary relation identification from scientific articles. We annotate our dataset by integrating automatic and human annotations, leveraging existing scientific knowledge resources. We develop a neural model as a strong baseline that extends previous state-of-the-art IE models to document-level IE. Analyzing the model performance shows a significant gap between human performance and current baselines, inviting the community to use our dataset as a challenge to develop document-level IE models. Our data and code are publicly available at https://github.com/allenai/SciREX .

Soft Gazetteers for Low-Resource Named Entity Recognition [Website][PDF]
Shruti Rijhwani, Shuyan Zhou, Graham Neubig, and Jaime Carbonell 21:00–22:00

Traditional named entity recognition models use gazetteers (lists of entities) as features to improve performance. Although modern neural network models do not require such hand-crafted features for strong performance, recent work has demonstrated their utility for named entity recognition on English data. However, designing such features for low-resource languages is challenging, because exhaustive entity gazetteers do not exist in these languages. To address this problem, we propose a method of "soft gazetteers" that incorporates ubiquitously available information from English knowledge bases, such as Wikipedia, into neural named entity recognition models through cross-lingual entity linking. Our experiments on four low-resource languages show an average improvement of 4 points in F1 score.

Sources of Transfer in Multilingual Named Entity Recognition [Website][PDF]
David Mueller, Nicholas Andrews, and Mark Dredze 21:00–22:00

Named-entities are inherently multilingual, and annotations in any given language may be limited. This motivates us to consider *polyglot* named-entity recognition (NER), where one model is trained using annotated data drawn from more than one language. However, a straightforward implementation of this simple idea does not always work in practice: naive training of NER models using annotated data drawn from multiple languages consistently underperforms models trained on monolingual data alone, despite having access to more training data. The starting point of this paper is a simple solution to this problem, in which polyglot models are *fine-tuned* on monolingual data to consistently and significantly outperform their monolingual counterparts. To explain this phenomena, we explore the sources of multilingual transfer in polyglot NER models and examine the weight structure of polyglot models compared to their monolingual counterparts. We find that polyglot models efficiently share many parameters across languages and that fine-tuning may utilize a large number of those parameters.

TXtract: Taxonomy-Aware Knowledge Extraction for Thousands of Product Categories [Website][PDF]
Giannis Karamanolakis, Jun Ma, and Xin Luna Dong 21:00–22:00

Extracting structured knowledge from product profiles is crucial for various applications in e-Commerce. State-of-the-art approaches for knowledge extraction were each designed for a single category of product, and thus do not apply to real-life e-Commerce scenarios, which often contain thousands of diverse categories. This paper proposes TXtract, a taxonomy-aware knowledge extraction model that applies to thousands of product categories organized in a hierarchical taxonomy. Through category conditional self-attention and multi-task learning, our approach is both scalable, as it trains a single model for thousands of categories, and effective, as it extracts category-specific attribute values. Experiments on products from a taxonomy with 4,000 categories show that TXtract outperforms state-of-the-art approaches by up to 10% in F1 and 15% in coverage across all categories.

TriggerNER: Learning with Entity Triggers as Explanations for Named Entity Recognition [Website][PDF]
Bill Yuchen Lin, Dong-Ho Lee, Ming Shen, Ryan Moreno, Xiao Huang, Prashant Shiralkar, and Xiang Ren 21:00–22:00

Training neural models for named entity recognition (NER) in a new domain often requires additional human annotations (e.g., tens of thousands of labeled instances) that are usually expensive and time-consuming to collect. Thus, a crucial research question is how to obtain supervision in a cost-effective way. In this paper, we introduce "entity triggers," an effective proxy of human explanations for facilitating label-efficient learning of NER models. An entity trigger is defined as a group of words in a sentence that helps to explain why humans would recognize an entity in the sentence. We crowd-sourced 14k entity triggers for two well-studied NER datasets. Our proposed model, Trigger Matching Network, jointly learns trigger representations and soft matching module with self-attention such that can generalize to unseen sentences easily for tagging. Our framework is significantly more cost-effective than the traditional neural NER frameworks. Experiments show that using only 20% of the trigger-annotated sentences results in a comparable performance as using 70% of conventional annotated sentences.

Session 15B: Machine Translation-18

Balancing Training for Multilingual Neural Machine Translation [Website][PDF]
Xinyi Wang, Yulia Tsvetkov, and Graham Neubig 21:00–22:00

When training multilingual machine translation (MT) models that can translate to/from multiple languages, we are faced with imbalanced training sets: some languages have much more training data than others. Standard practice is to up-sample less resourced languages to increase representation, and the degree of up-sampling has a large effect on the overall performance. In this paper, we propose a method that instead automatically learns how to weight training data through a data scorer that is optimized to maximize performance on all test languages. Experiments on two sets of languages under both one-to-many and many-to-one MT settings show our method not only consistently outperforms heuristic baselines in terms of average performance, but also offers flexible control over the performance of which languages are optimized.

HAT: Hardware-Aware Transformers for Efficient Natural Language Processing [Website][PDF]
Hanrui Wang, Zhanghao Wu, Zhijian Liu, Han Cai, Ligeng Zhu, Chuang Gan, and Song Han 21:00–22:00

Transformers are ubiquitous in Natural Language Processing (NLP) tasks, but they are difficult to be deployed on hardware due to the intensive computation. To enable low-latency inference on resource-constrained hardware platforms, we propose to design Hardware-Aware Transformers (HAT) with neural architecture search. We first construct a large design space with arbitrary encoder-decoder attention and heterogeneous layers. Then we train a SuperTransformer that covers all candidates in the design space, and efficiently produces many SubTransformers with weight sharing. Finally, we perform an evolutionary search with a hardware latency constraint to find a specialized SubTransformer dedicated to run fast on the target hardware. Extensive experiments on four machine translation tasks demonstrate that HAT can discover efficient models for different hardware (CPU, GPU, IoT device). When running WMT'14 translation task on Raspberry Pi-4, HAT can achieve 3× speedup, 3.7× smaller size over baseline Transformer; 2.7× speedup, 3.6× smaller size over Evolved Transformer with 12,041× less search cost and no performance loss. HAT is open-sourced at https://github.com/mit-han-lab/hardware-aware-transformers.

Parallel Corpus Filtering via Pre-trained Language Models [Website][PDF]
Boliang Zhang, Ajay Nagesh, and Kevin Knight 21:00–22:00

Web-crawled data provides a good source of parallel corpora for training machine translation models. It is automatically obtained, but extremely noisy, and recent work shows that neural machine translation systems are more sensitive to noise than traditional statistical machine translation methods. In this paper, we propose a novel approach to filter out noisy sentence pairs from web-crawled corpora via pre-trained language models. We measure sentence parallelism by leveraging the multilingual capability of BERT and use the Generative Pre-training (GPT) language model as a domain filter to balance data domains. We evaluate the proposed method on the WMT 2018 Parallel Corpus Filtering shared task, and on our own web-crawled Japanese-Chinese parallel corpus. Our method significantly outperforms baselines and achieves a new state-of-the-art. In an unsupervised setting, our method achieves comparable performance to the top-1 supervised method. We also evaluate on a web-crawled Japanese-Chinese parallel corpus that we make publicly available.

Regularized Context Gates on Transformer for Machine Translation [Website][PDF]
Xintong Li, Lemao Liu, Rui Wang, Guoping Huang, and Max Meng 21:00–22:00

Context gates are effective to control the contributions from the source and target contexts in the recurrent neural network (RNN) based neural machine translation (NMT). However, it is challenging to extend them into the advanced Transformer architecture, which is more complicated than RNN. This paper first provides a method to identify source and target contexts and then introduce a gate mechanism to control the source and target contributions in Transformer. In addition, to further reduce the bias problem in the gate mechanism, this paper proposes a regularization method to guide the learning of the gates with supervision automatically generated using pointwise mutual information. Extensive experiments on 4 translation datasets demonstrate that the proposed model obtains an averaged gain of 1.0 BLEU score over a strong Transformer baseline.

Translationese as a Language in "Multilingual" NMT [Website][PDF]
Parker Riley, Isaac Caswell, Markus Freitag, and David Grangier 21:00–22:00

Machine translation has an undesirable propensity to produce "translationese" artifacts, which can lead to higher BLEU scores while being liked less by human raters. Motivated by this, we model translationese and original (i.e. natural) text as separate languages in a multilingual model, and pose the question: can we perform zero-shot translation between original source text and original target text? There is no data with original source and original target, so we train a sentence-level classifier to distinguish translationese from original target text, and use this classifier to tag the training data for an NMT model. Using this technique we bias the model to produce more natural outputs at test time, yielding gains in human evaluation scores on both accuracy and fluency. Additionally, we demonstrate that it is possible to bias the model to produce translationese and game the BLEU score, increasing it while decreasing human-rated quality. We analyze these outputs using metrics measuring the degree of translationese, and present an analysis of the volatility of heuristic-based train-data tagging.

Session 15B: NLP Applications-12

A Multi-Perspective Architecture for Semantic Code Search [Website][PDF]
Rajarshi Haldar, Lingfei Wu, JinJun Xiong, and Julia Hockenmaier 21:00–22:00

The ability to match pieces of code to their corresponding natural language descriptions and vice versa is fundamental for natural language search interfaces to software repositories. In this paper, we propose a novel multi-perspective cross-lingual neural framework for code–text matching, inspired in part by a previous model for monolingual text-to-text matching, to capture both global and local similarities. Our experiments on the CoNaLa dataset show that our proposed model yields better performance on this cross-lingual text-to-code matching task than previous approaches that map code and text to a single joint embedding space.

Automated Topical Component Extraction Using Neural Network Attention Scores from Source-based Essay Scoring [Website][PDF]
Haoran Zhang and Diane Litman 21:00–22:00

While automated essay scoring (AES) can reliably grade essays at scale, automated writing evaluation (AWE) additionally provides formative feedback to guide essay revision. However, a neural AES typically does not provide useful feature representations for supporting AWE. This paper presents a method for linking AWE and neural AES, by extracting Topical Components (TCs) representing evidence from a source text using the intermediate output of attention layers. We evaluate performance using a feature-based AES requiring TCs. Results show that performance is comparable whether using automatically or manually constructed TCs for 1) representing essays as rubric-based features, 2) grading essays.

Clinical Concept Linking with Contextualized Neural Representations [Website][PDF]
Elliot Schumacher, Andriy Mulyar, and Mark Dredze 21:00–22:00

In traditional approaches to entity linking, linking decisions are based on three sources of information – the similarity of the mention string to an entity's name, the similarity of the context of the document to the entity, and broader information about the knowledge base (KB). In some domains, there is little contextual information present in the KB and thus we rely more heavily on mention string similarity. We consider one example of this, concept linking, which seeks to link mentions of medical concepts to a medical concept ontology. We propose an approach to concept linking that leverages recent work in contextualized neural models, such as ELMo (Peters et al. 2018), which create a token representation that integrates the surrounding context of the mention and concept name. We find a neural ranking approach paired with contextualized embeddings provides gains over a competitive baseline (Leaman et al. 2013). Additionally, we find that a pre-training step using synonyms from the ontology offers a useful initialization for the ranker.

DeSePtion: Dual Sequence Prediction and Adversarial Examples for Improved Fact-Checking [Website][PDF]
Christopher Hidey, Tuhin Chakrabarty, Tariq Alhindi, Siddharth Varia, Kriste Krstovski, Mona Diab, and Smaranda Muresan 21:00–22:00

The increased focus on misinformation has spurred development of data and systems for detecting the veracity of a claim as well as retrieving authoritative evidence. The Fact Extraction and VERification (FEVER) dataset provides such a resource for evaluating endto- end fact-checking, requiring retrieval of evidence from Wikipedia to validate a veracity prediction. We show that current systems for FEVER are vulnerable to three categories of realistic challenges for fact-checking — multiple propositions, temporal reasoning, and ambiguity and lexical variation — and introduce a resource with these types of claims. Then we present a system designed to be resilient to these "attacks" using multiple pointer networks for document selection and jointly modeling a sequence of evidence sentences and veracity relation predictions. We find that in handling these attacks we obtain state-of-the-art results on FEVER, largely due to improved evidence retrieval.

Let Me Choose: From Verbal Context to Font Selection [Website][PDF]
Amirreza Shirani, Franck Dernoncourt, Jose Echevarria, Paul Asente, Nedim Lipka, and Thamar Solorio 21:00–22:00

In this paper, we aim to learn associations between visual attributes of fonts and the verbal context of the texts they are typically applied to. Compared to related work leveraging the surrounding visual context, we choose to focus only on the input text, which can enable new applications for which the text is the only visual element in the document. We introduce a new dataset, containing examples of different topics in social media posts and ads, labeled through crowd-sourcing. Due to the subjective nature of the task, multiple fonts might be perceived as acceptable for an input text, which makes this problem challenging. To this end, we investigate different end-to-end models to learn label distributions on crowd-sourced data, to capture inter-subjectivity across all annotations.

[TACL] Machine Learning Driven Language Assessment [Website][PDF]
Burr Settles, Masato Hagiwara, and Geoffrey T. LaFlair 21:00–22:00

We describe a method for rapidly creating language proficiency assessments, and provide experimental evidence that such tests can be valid, reliable, and secure. Our approach is the first to use machine learning and natural language processing to induce proficiency scales based on a given standard, and then use linguistic models to estimate item difficulty directly for computer-adaptive testing. This alleviates the need for expensive pilot testing with human subjects. We used these methods to develop an online proficiency exam called the Duolingo English Test, and demonstrate that its scores align significantly with other high-stakes English assessments. Furthermore, our approach produces test scores that are highly reliable, while generating item banks large enough to satisfy security requirements.

Multi-Label and Multilingual News Framing Analysis [Website][PDF]
Afra Feyza Akyürek, Lei Guo, Randa Elanwar, Prakash Ishwar, Margrit Betke, and Derry Tanti Wijaya
21:00–22:00

News framing refers to the practice in which aspects of specific issues are highlighted in the news to promote a particular interpretation. In NLP, although recent works have studied framing in English news, few have studied how the analysis can be extended to other languages and in a multi-label setting. In this work, we explore multilingual transfer learning to detect multiple frames from just the news headline in a genuinely low-resource context where there are few/no frame annotations in the target language. We propose a novel method that can leverage elementary resources consisting of a dictionary and few annotations to detect frames in the target language. Our method performs comparably or better than translating the entire target language headline to the source language for which we have annotated data. This work opens up an exciting new capability of scaling up frame analysis to many languages, even those without existing translation technologies. Lastly, we apply our method to detect frames on the issue of U.S. gun violence in multiple languages and obtain exciting insights on the relationship between different frames of the same problem across different countries with different languages.

Predicting Performance for Natural Language Processing Tasks [Website][PDF]
Mengzhou Xia, Antonios Anastasopoulos, Ruochen Xu, Yiming Yang, and Graham Neubig 21:00–22:00

Given the complexity of combinations of tasks, languages, and domains in natural language processing (NLP) research, it is computationally prohibitive to exhaustively test newly proposed models on each possible experimental setting. In this work, we attempt to explore the possibility of gaining plausible judgments of how well an NLP model can perform under an experimental setting, *without actually training or testing the model*. To do so, we build regression models to predict the evaluation score of an NLP experiment given the experimental settings as input. Experimenting on~9 different NLP tasks, we find that our predictors can produce meaningful predictions over unseen languages and different modeling architectures, outperforming reasonable baselines as well as human experts. %we represent experimental settings using an array of features. Going further, we outline how our predictor can be used to find a small subset of representative experiments that should be run in order to obtain plausible predictions for all other experimental settings.[2]

ScriptWriter: Narrative-Guided Script Generation [Website][PDF]
Yutao Zhu, Ruihua Song, Zhicheng Dou, Jian-Yun NIE, and Jin Zhou 21:00–22:00

It is appealing to have a system that generates a story or scripts automatically from a storyline, even though this is still out of our reach. In dialogue systems, it would also be useful to drive dialogues by a dialogue plan. In this paper, we address a key problem involved in these applications - guiding a dialogue by a narrative. The proposed model ScriptWriter selects the best response among the candidates that fit the context as well as the given narrative. It keeps track of what in the narrative has been said and what is to be said. A narrative plays a different role than the context (i.e., previous utterances), which is generally used in current dialogue systems. Due to the unavailability of data for this new application, we construct a new large-scale data collection GraphMovie from a movie website where end-users can upload their narratives freely when watching a movie. Experimental results on the dataset show that our proposed approach based on narratives significantly outperforms the baselines that simply use the narrative as a kind of context.

Should All Cross-Lingual Embeddings Speak English? [Website][PDF]
Antonios Anastasopoulos and Graham Neubig 21:00–22:00

Most of recent work in cross-lingual word embeddings is severely Anglocentric. The vast majority of lexicon induction evaluation dictionaries are between English and another language, and the English embedding space is selected by default as the hub when learning in a multilingual setting. With this work, however, we challenge these practices. First, we show that the choice of hub language can significantly impact downstream lexicon induction zero-shot POS tagging performance. Second, we both expand a standard English-centered evaluation dictionary collection to include all language pairs using triangulation, and create new dictionaries for under-represented languages. Evaluating established methods over all these language pairs sheds light into their suitability for aligning embeddings from distant languages and presents new challenges for the field. Finally, in our analysis we identify general guidelines for strong cross-lingual embedding baselines, that extend to language pairs that do not include English.

Smart To-Do: Automatic Generation of To-Do Items from Emails [Website][PDF]
Sudipto Mukherjee, Subhabrata Mukherjee, Marcello Hasegawa, Ahmed Hassan Awadallah, and Ryen White 21:00–22:00

Intelligent features in email service applications aim to increase productivity by helping people organize their folders, compose their emails and respond to pending tasks. In this work, we explore a new application, Smart-To-Do, that helps users with task management over emails. We introduce a new task and dataset for automatically generating To-Do items from emails where the sender has promised to perform an action. We design a two-stage process leveraging recent advances in neural text generation and sequence-to-sequence learning, obtaining BLEU and ROUGE scores of 0.23 and 0.63 for this task. To the best of our knowledge, this is the first work to address the problem of composing To-Do items from emails.

[2] Code, data and logs are publicly available at `https://github.com/xiamengzhou/NLPerf`.

Session 15B: Phonology, Morphology and Word Segmentation-7

A Multitask Learning Approach for Diacritic Restoration [Website][PDF]
Sawsan Alqahtani, Ajay Mishra, and Mona Diab 21:00–22:00

In many languages like Arabic, diacritics are used to specify pronunciations as well as meanings. Such diacritics are often omitted in written text, increasing the number of possible pronunciations and meanings for a word. This results in a more ambiguous text making computational processing on such text more difficult. Diacritic restoration is the task of restoring missing diacritics in the written text. Most state-of-the-art diacritic restoration models are built on character level information which helps generalize the model to unseen data, but presumably lose useful information at the word level. Thus, to compensate for this loss, we investigate the use of multi-task learning to jointly optimize diacritic restoration with related NLP problems namely word segmentation, part-of-speech tagging, and syntactic diacritization. We use Arabic as a case study since it has sufficient data resources for tasks that we consider in our joint modeling. Our joint models significantly outperform the baselines and are comparable to the state-of-the-art models that are more complex relying on morphological analyzers and/or a lot more data (e.g. dialectal data).

Frugal Paradigm Completion [Website][PDF]
Alexander Erdmann, Tom Kenter, Markus Becker, and Christian Schallhart 21:00–22:00

Lexica distinguishing all morphologically related forms of each lexeme are crucial to many language technologies, yet building them is expensive. We propose a frugal paradigm completion approach that predicts all related forms in a morphological paradigm from as few manually provided forms as possible. It induces typological information during training which it uses to determine the best sources at test time. We evaluate our language-agnostic approach on 7 diverse languages. Compared to popular alternative approaches, ours reduces manual labor by 16-63% and is the most robust to typological variation.

Improving Chinese Word Segmentation with Wordhood Memory Networks [Website][PDF]
Yuanhe Tian, Yan Song, Fei Xia, Tong Zhang, and Yonggang Wang 21:00–22:00

Contextual features always play an important role in Chinese word segmentation (CWS). Wordhood information, being one of the contextual features, is proved to be useful in many conventional character-based segmenters. However, this feature receives less attention in recent neural models and it is also challenging to design a framework that can properly integrate wordhood information from different wordhood measures to existing neural frameworks. In this paper, we therefore propose a neural framework, WMSeg, which uses memory networks to incorporate wordhood information with several popular encoder-decoder combinations for CWS. Experimental results on five benchmark datasets indicate the memory mechanism successfully models wordhood information for neural segmenters and helps WMSeg achieve state-of-the-art performance on all those datasets. Further experiments and analyses also demonstrate the robustness of our proposed framework with respect to different wordhood measures and the efficiency of wordhood information in cross-domain experiments.

Joint Chinese Word Segmentation and Part-of-speech Tagging via Two-way Attentions of Auto-analyzed Knowledge [Website][PDF]
Yuanhe Tian, Yan Song, Xiang Ao, Fei Xia, Xiaojun Quan, Tong Zhang, and Yonggang Wang 21:00–22:00

Chinese word segmentation (CWS) and part-of-speech (POS) tagging are important fundamental tasks for Chinese language processing, where joint learning of them is an effective one-step solution for both tasks. Previous studies for joint CWS and POS tagging mainly follow the character-based tagging paradigm with introducing contextual information such as n-gram features or sentential representations from recurrent neural models. However, for many cases, the joint tagging needs not only modeling from context features but also knowledge attached to them (e.g., syntactic relations among words); limited efforts have been made by existing research to meet such needs. In this paper, we propose a neural model named TwASP for joint CWS and POS tagging following the character-based sequence labeling paradigm, where a two-way attention mechanism is used to incorporate both context feature and their corresponding syntactic knowledge for each input character. Particularly, we use existing language processing toolkits to obtain the auto-analyzed syntactic knowledge for the context, and the proposed attention module can learn and benefit from them although their quality may not be perfect. Our experiments illustrate the effectiveness of the two-way attentions for joint CWS and POS tagging, where state-of-the-art performance is achieved on five benchmark datasets.

Joint Diacritization, Lemmatization, Normalization, and Fine-Grained Morphological Tagging [Website][PDF]
Nasser Zalmout and Nizar Habash 21:00–22:00

The written forms of Semitic languages are both highly ambiguous and morphologically rich: a word can have multiple interpretations and is one of many inflected forms of the same concept or lemma. This is further exacerbated for dialectal content, which is more prone to noise and lacks a standard orthography. The morphological features can be lexicalized, like lemmas and diacritized forms, or non-lexicalized, like gender, number, and part-of-speech tags, among others. Joint modeling of the lexicalized and non-lexicalized features can identify more intricate morphological patterns, which provide better context modeling, and further disambiguate ambiguous lexical choices. However, the different modeling granularity can make joint modeling more difficult. Our approach models the different features jointly, whether lexicalized (on the character-level), or non-lexicalized (on the word-level). We use Arabic as a test case, and achieve state-of-the-art results for Modern Standard Arabic with 20% relative error reduction, and Egyptian Arabic with 11% relative error reduction.

Phonetic and Visual Priors for Decipherment of Informal Romanization [Website][PDF]
Maria Ryskina, Matthew R. Gormley, and Taylor Berg-Kirkpatrick 21:00–22:00

Informal romanization is an idiosyncratic process used by humans in informal digital communication to encode non-Latin script languages into Latin character sets found on common keyboards. Character substitution choices differ between users but have been shown to be governed by the same main principles observed across a variety of languages—namely, character pairs are often associated through phonetic or visual similarity. We propose a noisy-channel WFST cascade model for deciphering the original non-Latin script from observed romanized text in an unsupervised fashion. We train our model directly on romanized data from two languages: Egyptian Arabic and Russian. We demonstrate that adding inductive bias through phonetic and visual priors on character mappings substantially improves the model's performance on both languages, yielding results much closer to the supervised skyline. Finally, we introduce a new dataset of romanized Russian, collected from a Russian social network website and partially annotated for our experiments.

Unsupervised Morphological Paradigm Completion [Website] [PDF]
Huiming Jin, Liwei Cai, Yihui Peng, Chen Xia, Arya McCarthy, and Katharina Kann 21:00–22:00

We propose the task of unsupervised morphological paradigm completion. Given only raw text and a lemma list, the task consists of generating the morphological paradigms, i.e., all inflected forms, of the lemmas. From a natural language processing (NLP) perspective, this is a challenging unsupervised task, and high-performing systems have the potential to improve tools for low-resource languages or to assist linguistic annotators. From a cognitive science perspective, this can shed light on how children acquire morphological knowledge. We further introduce a system for the task, which generates morphological paradigms via the following steps: (i) EDIT TREE retrieval, (ii) additional lemma retrieval, (iii) paradigm size discovery, and (iv) inflection generation. We perform an evaluation on 14 typologically diverse languages. Our system outperforms trivial baselines with ease and, for some languages, even obtains a higher accuracy than minimally supervised systems.

Session 15B Semantics: Sentence Level-11

Beyond Possession Existence: Duration and Co-Possession [Website][PDF]
Dhivya Chinnappa, Srikala Murugan, and Eduardo Blanco 21:00–22:00

This paper introduces two tasks: determining (a) the duration of possession relations and (b) co-possessions, i.e., whether multiple possessors possess a possessee at the same time. We present new annotations on top of corpora annotating possession existence and experimental results. Regarding possession duration, we derive the time spans we work with empirically from annotations indicating lower and upper bounds. Regarding co-possessions, we use a binary label. Cohen's kappa coefficients indicate substantial agreement, and experimental results show that text is more useful than the image for solving these tasks.

Controlled Crowdsourcing for High-Quality QA-SRL Annotation [Website][PDF]
Paul Roit, Ayal Klein, Daniela Stepanov, Jonathan Mamou, Julian Michael, Gabriel Stanovsky, Luke Zettlemoyer, and Ido Dagan 21:00–22:00

Question-answer driven Semantic Role Labeling (QA-SRL) was proposed as an attractive open and natural flavour of SRL, potentially attainable from laymen. Recently, a large-scale crowdsourced QA-SRL corpus and a trained parser were released. Trying to replicate the QA-SRL annotation for new texts, we found that the resulting annotations were lacking in quality, particularly in coverage, making them insufficient for further research and evaluation. In this paper, we present an improved crowdsourcing protocol for complex semantic annotation, involving worker selection and training, and a data consolidation phase. Applying this protocol to QA-SRL yielded high-quality annotation with drastically higher coverage, producing a new gold evaluation dataset. We believe that our annotation protocol and gold standard will facilitate future replicable research of natural semantic annotations.

Don't Stop Pretraining: Adapt Language Models to Domains and Tasks [Website][PDF]
Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith 21:00–22:00

Language models pretrained on text from a wide variety of sources form the foundation of today's NLP. In light of the success of these broad-coverage models, we investigate whether it is still helpful to tailor a pretrained model to the domain of a target task. We present a study across four domains (biomedical and computer science publications, news, and reviews) and eight classification tasks, showing that a second phase of pretraining in-domain (domain-adaptive pretraining) leads to performance gains, under both high- and low-resource settings. Moreover, adapting to the task's unlabeled data (task-adaptive pretraining) improves performance even after domain-adaptive pretraining. Finally, we show that adapting to a task corpus augmented using simple data selection strategies is an effective alternative, especially when resources for domain-adaptive pretraining might be unavailable. Overall, we consistently find that multi-phase adaptive pretraining offers large gains in task performance.

Structured Tuning for Semantic Role Labeling [Website][PDF]
Tao Li, Parth Anand Jawale, Martha Palmer, and Vivek Srikumar 21:00–22:00

Recent neural network-driven semantic role labeling (SRL) systems have shown impressive improvements in F1 scores. These improvements are due to expressive input representations, which, at least at the surface, are orthogonal to knowledge-rich constrained decoding mechanisms that helped linear SRL models. Introducing the benefits of structure to inform neural models presents a methodological challenge. In this paper, we present a structured tuning framework to improve models using softened constraints only at training time. Our framework leverages the expressiveness of neural networks and provides supervision with structured loss components. We start with a strong baseline (RoBERTa) to validate the impact of our approach, and show that our framework outperforms the baseline by learning to comply with declarative constraints. Additionally, our experiments with smaller training sizes show that we can achieve consistent improvements under low-resource scenarios.

Universal Decompositional Semantic Parsing [Website][PDF]
Elias Stengel-Eskin, Aaron Steven White, Sheng Zhang, and Benjamin Van Durme 21:00–22:00

We introduce a transductive model for parsing into Universal Decompositional Semantics (UDS) representations, which jointly learns to map natural language utterances into UDS graph structures and annotate the graph with decompositional semantic attribute scores. We also introduce a strong pipeline model for parsing into the UDS graph structure, and show that our transductive parser performs comparably while additionally performing attribute prediction. By analyzing the attribute prediction errors, we find the model captures natural relationships between attribute groups.

Session 15B: Student Research Workshop

#NotAWhore! A Computational Linguistic Perspective of Rape Culture and Victimization on Social Media [Website][PDF]
Ashima Suvarna and Grusha Bhalla 21:00–22:00

The recent surge in online forums and movements supporting sexual assault survivors has led to the emergence of a 'virtual bubble' where survivors can recount their stories. However, this also makes the survivors vulnerable to bullying, trolling and victim blaming. Specifically, victim blaming has been shown to have acute psychological effects on the survivors and further discourage formal reporting of such crimes. Therefore, it is important to devise computationally relevant methods to identify and prevent victim blaming to protect the victims. In our work, we discuss the drastic effects of victim blaming through a short case study and then propose a single step transfer-learning based classification method to identify victim blaming language on Twitter. Finally, we compare the performance of our proposed model against various deep learning and machine learning models on a manually annotated domain-specific dataset.

Crossing the Line: Where do Demographic Variables Fit into Humor Detection? [Website][PDF]
J. A. Meaney 21:00–22:00

Recent humor classification shared tasks have struggled with two issues: either the data comprises a highly constrained genre of humor which does not broadly represent humor, or the data is so indiscriminate that the inter-annotator agreement on its humor content is drastically low. These tasks typically average over all annotators' judgments, in spite of the fact that humor is a highly subjective phenomenon. We argue that demographic factors influence whether a text is perceived as humorous or not. We propose the addition of demographic information about the humor annotators in order to bin ratings more sensibly. We also suggest the addition of an 'offensive' label to distinguish between different generations, in terms of humor. This would allow for more nuanced shared tasks and could lead to better performance on downstream tasks, such as content moderation.

Effectively Aligning and Filtering Parallel Corpora under Sparse Data Conditions [Website][PDF]
Steinþór Steingrímsson, Hrafn Loftsson, and Andy Way 21:00–22:00

Parallel corpora are key to developing good machine translation systems. However, abundant parallel data are hard to come by, especially for languages with a low number of speakers. When rich morphology exacerbates the data sparsity problem, it is imperative to have accurate alignment and filtering methods that can help make the most of what is available by maximising the number of correctly translated segments in a corpus and minimising noise by removing incorrect translations and segments containing extraneous data. This paper sets out a research plan for improving alignment and filtering methods for parallel texts in low-resource settings. We propose an effective unsupervised alignment method to tackle the alignment problem. Moreover, we propose a strategy to supplement state-of-the-art models with automatically extracted information using basic NLP tools to effectively handle rich morphology.

Exploring the Role of Context to Distinguish Rhetorical and Information-Seeking Questions [Website][PDF]
Yuan Zhuang and Ellen Riloff 21:00–22:00

Social media posts often contain questions, but many of the questions are rhetorical and do not seek information. Our work studies the problem of distinguishing rhetorical and information-seeking questions on Twitter. Most work has focused on features of the question itself, but we hypothesize that the prior context plays a role too. This paper introduces a new dataset containing questions in tweets paired with their prior tweets to provide context. We create classification models to assess the difficulty of distinguishing rhetorical and information-seeking questions, and experiment with different properties of the prior context. Our results show that the prior tweet and topic features can improve performance on this task.

Session 15B Syntax: Tagging, Chunking and Parsing-6

[TACL] Deep Contextualized Self-training for Low Resource Dependency Parsing [Website][PDF]
Guy Rotman and Roi Reichart 21:00–22:00

Neural dependency parsing has proven very effective, achieving state-of-the-art results on numerous domains and languages. Unfortunately, it requires large amounts of labeled data, that is costly and laborious to create. In this paper we propose a self-training algorithm that alleviates this annotation bottleneck by training a parser on its own output. Our Deep Contextualized Self-training (DCST) algorithm utilizes representation models trained on sequence labeling tasks that are derived from the parser's output when applied to unlabeled data and integrates these models with the base parser through a gating mechanism. We conduct experiments across multiple languages, both in low resource in-domain and in cross-domain setups and demonstrate that DCST substantially outperforms traditional self-training as well as recent semi-supervised training methods.

Extracting Headless MWEs from Dependency Parse Trees: Parsing, Tagging, and Joint Modeling Approaches [Website][PDF]
Tianze Shi and Lillian Lee 21:00–22:00

An interesting and frequent type of multi-word expression (MWE) is the headless MWE, for which there are no true internal syntactic dominance relations; examples include many named entities ("Wells Fargo") and dates ("July 5, 2020") as well as certain productive constructions ("blow for blow", "day after day"). Despite their special status and prevalence, current dependency-annotation schemes require treating such flat structures as if they had internal syntactic heads, and most current parsers handle them in the same fashion as headed constructions. Meanwhile, outside the context of parsing, taggers are typically used for identifying MWEs, but taggers might benefit from structural information. We empirically compare these two common strategies—parsing and tagging—for predicting flat MWEs. Additionally, we propose an efficient joint decoding algorithm that combines scores from both strategies. Experimental results on the MWE-Aware English Dependency Corpus and on six non-English dependency treebanks with frequent flat structures show that: (1) tagging is more accurate than parsing for identifying flat-structure MWEs, (2) our joint decoder reconciles the two different views and, for non-BERT features, leads to higher accuracies, and (3) most of the gains result from feature sharing between the parsers and taggers.

Revisiting Higher-Order Dependency Parsers [Website][PDF]
Erick Fonseca and André F. T. Martins 21:00–22:00

Neural encoders have allowed dependency parsers to shift from higher-order structured models to simpler first-order ones, making decoding faster and still achieving better accuracy than non-neural parsers. This has led to a belief that neural encoders can implicitly encode structural constraints, such as siblings and grandparents in a tree. We tested this hypothesis and found that neural parsers may benefit from higher-order features, even when employing a powerful pre-trained encoder, such as BERT. While the gains of higher-order features are small in the presence of a powerful encoder, they are consistent for long-range dependencies and long sentences. In particular, higher-order models are more accurate on full sentence parses and on the exact match of modifier lists, indicating that they deal better with larger, more complex structures.

SeqVAT: Virtual Adversarial Training for Semi-Supervised Sequence Labeling [Website][PDF]
Luoxin Chen, Weitong Ruan, Xinyue Liu, and Jianhua Lu 21:00–22:00

Virtual adversarial training (VAT) is a powerful technique to improve model robustness in both supervised and semi-supervised settings. It is effective and can be easily adopted on lots of image classification and text classification tasks. However, its benefits to sequence labeling tasks such as named entity recognition (NER) have not been shown as significant, mostly, because the previous approach can not combine VAT with the conditional random field (CRF). CRF can significantly boost accuracy for sequence models by putting constraints on label transitions, which makes it an essential component in most state-of-the-art sequence labeling model architectures. In this paper, we propose SeqVAT, a method which naturally applies VAT to sequence labeling models with CRF. Empirical studies show that SeqVAT not only significantly improves the sequence labeling performance over baselines under supervised settings, but also outperforms state-of-the-art approaches under semi-supervised settings.

Treebank Embedding Vectors for Out-of-domain Dependency Parsing [Website][PDF]
Joachim Wagner, James Barry, and Jennifer Foster 21:00–22:00

A recent advance in monolingual dependency parsing is the idea of a treebank embedding vector, which allows all treebanks for a particular language to be used as training data while at the same time allowing the model to prefer training data from one treebank over others and to select the preferred treebank at test time. We build on this idea by 1) introducing a method to predict a treebank vector for sentences that do not come from a treebank used in training, and 2) exploring what happens when we move away from predefined treebank embedding vectors during test time and instead devise tailored interpolations. We show that 1) there are interpolated vectors that are superior to the predefined ones, and 2) treebank vectors can be predicted with sufficient accuracy, for nine out of ten test languages, to match the performance of an oracle approach that knows the most suitable predefined treebank embedding for the test set.

Session 15B: Theme-6

Automated Evaluation of Writing — 50 Years and Counting [Website][PDF]
Beata Beigman Klebanov and Nitin Madnani 21:00–22:00

In this theme paper, we focus on Automated Writing Evaluation (AWE), using Ellis Page's seminal 1966 paper to frame the presentation. We discuss some of the current frontiers in the field and offer some thoughts on the emergent uses of this technology.

Negated and Misprimed Probes for Pretrained Language Models: Birds Can Talk, But Cannot Fly [Website][PDF]
Nora Kassner and Hinrich Schütze 21:00–22:00

Building on Petroni et al. 2019, we propose two new probing tasks analyzing factual knowledge stored in Pretrained Language Models (PLMs). (1) Negation. We find that PLMs do not distinguish between negated ("Birds cannot [MASK]") and non-negated ("Birds can [MASK]") cloze questions. (2) Mispriming. Inspired by priming methods in human psychology, we add "misprimes" to cloze questions ("Talk? Birds can [MASK]"). We find that PLMs are easily distracted by misprimes. These results suggest that PLMs still have a long way to go to adequately learn human-like factual knowledge.

On Forgetting to Cite Older Papers: An Analysis of the ACL Anthology [Website][PDF]
Marcel Bollmann and Desmond Elliott 21:00–22:00

The field of natural language processing is experiencing a period of unprecedented growth, and with it a surge of published papers. This represents an opportunity for us to take stock of how we cite the work of other researchers, and whether this growth comes at the expense of "forgetting" about older literature. In this paper, we address this question through bibliographic analysis. By looking at the age of outgoing citations in papers published at selected ACL venues between 2010 and 2019, we find that there is indeed a tendency for recent papers to cite more recent work, but the rate at which papers older than 15 years are cited has remained relatively stable.

Returning the N to NLP: Towards Contextually Personalized Classification Models [Website][PDF]
Lucie Flek 21:00–22:00

Most NLP models today treat language as universal, even though socio- and psycholingustic research shows that the communicated message is influenced by the characteristics of the speaker as well as the target audience. This paper surveys the landscape of personalization in natural language processing and related fields, and offers a path forward to mitigate the decades of deviation of the NLP tools from sociolingustic findings, allowing to flexibly process the "natural" language of each user rather than enforcing a uniform NLP treatment. It outlines a possible direction to incorporate these aspects into neural NLP models by means of socially contextual personalization, and proposes to shift the focus of our evaluation strategies accordingly.

The Unstoppable Rise of Computational Linguistics in Deep Learning [Website][PDF]
James Henderson 21:00–22:00

In this paper, we trace the history of neural networks applied to natural language understanding tasks, and identify key contributions which the nature of language has made to the development of neural network architectures. We focus on the importance of variable binding and its instantiation in attention-based models, and argue that Transformer is not a sequence model but an induced-structure model. This perspective leads to predictions of the challenges facing research in deep learning architectures for natural language understanding.

To Boldly Query What No One Has Annotated Before? The Frontiers of Corpus Querying [Website][PDF]
Markus Gärtner and Kerstin Jung 21:00–22:00

Corpus query systems exist to address the multifarious information needs of any person interested in the content of annotated corpora. In this role they play an important part in making those resources usable for a wider audience. Over the past decades, several such query systems and languages have emerged, varying greatly in their expressiveness and technical details. This paper offers a broad overview of the history of corpora and corpus query tools. It focusses strongly on the query side and hints at exciting directions for future development.

To Test Machine Comprehension, Start by Defining Comprehension [Website][PDF]
Jesse Dunietz, Greg Burnham, Akash Bharadwaj, Owen Rambow, Jennifer Chu-Carroll, and Dave Ferrucci 21:00–22:00

Many tasks aim to measure machine reading comprehension (MRC), often focusing on question types presumed to be difficult. Rarely, however, do task designers start by considering what systems should in fact comprehend. In this paper we make two key contributions. First, we argue that existing approaches do not adequately define comprehension; they are too unsystematic about what content is tested. Second, we present a detailed definition of comprehension—a "Template of Understanding"—for a widely useful class of texts, namely short narratives. We then conduct an experiment that strongly suggests existing systems are not up to the task of narrative understanding as we define it.

Gender Gap in Natural Language Processing Research: Disparities in Authorship and Citations [Website][PDF]
Saif M. Mohammad 21:00–22:00

Disparities in authorship and citations across genders can have substantial adverse consequences not just on the dis-

advantaged gender, but also on the field of study as a whole. In this work, we examine female first author percentages and the citations to their papers in Natural Language Processing. We find that only about 29% of first authors are female and only about 25% of last authors are female. Notably, this percentage has not improved since the mid 2000s. We also show that, on average, female first authors are cited less than male first authors, even when controlling for experience and area of research. We hope that recording citation and participation gaps across demographic groups will improve awareness of gender gaps and encourage more inclusiveness and fairness in research.

Demo Session 5C

Time: *21:30–22:15*

What's The Latest? A Question-driven News Chatbot [Website] [PDF]
Philippe Laban, John Canny, and Marti A. Hearst

This work describes an automatic news chatbot that draws content from a diverse set of news articles and creates conversations with a user about the news. Key components of the system include the automatic organization of news articles into topical chatrooms, integration of automatically generated questions into the conversation, and a novel method for choosing which questions to present which avoids repetitive suggestions. We describe the algorithmic framework and present the results of a usability study that shows that news readers using the system successfully engage in multi-turn conversations about specific news stories.

Torch-Struct: Deep Structured Prediction Library [Website] [PDF]
Alexander Rush

The literature on structured prediction for NLP describes a rich collection of distributions and algorithms over sequences, segmentations, alignments, and trees; however, these algorithms are difficult to utilize in deep learning frameworks. We introduce Torch-Struct, a library for structured prediction designed to take advantage of and integrate with vectorized, auto-differentiation based frameworks. Torch-Struct includes a broad collection of probabilistic structures accessed through a simple and flexible distribution-based API that connects to any deep learning model. The library utilizes batched, vectorized operations and exploits auto-differentiation to produce readable, fast, and testable code. Internally, we also include a number of general-purpose optimizations to provide cross-algorithm efficiency. Experiments show significant performance gains over fast baselines and case-studies demonstrate the benefits of the library. Torch-Struct is available at https://github.com/harvardnlp/pytorch-struct.

6 Workshops

Sunday, July 5

15:00–1:00 (+1)	**W1: WiNLP: The Fourth Widening NLP Workshop**	[Web]

Thursday, July 9

8:00–9:45	**W10: RepL4NLP (Session 1): The 5th Workshop on Representation Learning for NLP**	[Web]
11:00–20:15	**W5: FEVER: The Third workshop on Fact Extraction and VERification**	[Web]
13:00–20:00	**W3: NLP4ConvAI: NLP for Conversational AI workshop**	[Web]
13:00–21:00	**W7: FLP: The 2nd Workshop on Figurative Language Processing**	[Web]
14:00–17:00	**W6: IWPT: The 16th International Conference on Parsing Technologies**	[Web]
14:00–17:00	**W10: RepL4NLP (Session 2): The 5th Workshop on Representation Learning for NLP**	[Web]
15:00–0:15 (+1)	**W8: NUSE: The 1st Joint Workshop on Narrative Understanding, Storylines, and Events**	[Web]
15:30–1:00 (+1)	**W4: BioNLP 2020: Workshop on Biomedical Natural Language Processing**	[Web]
16:00–0:45 (+1)	**W9: ALVR: Workshop on Advances in Language and Vision Research**	[Web]
16:00–20:30	**W2: IWSLT: The 17th International Conference on Spoken Language Translation**	[Web]

Friday, July 10

0:00–3:00	**W10: RepL4NLP (Session 3): The 5th Workshop on Representation Learning for NLP**	[Web]
0:30–4:30	**W20: NLPCovid: NLP COVID-19 Workshop**	[Web]
12:30–21:15	**W16: ECNLP: The Third Workshop on e-Commerce and NLP**	[Web]
13:00–20:15	**W13: BEA: The 15th Workshop on Innovative Use of NLP for Building Educational Applications**	[Web]
13:00–20:45	**W19: Challenge-HML: The Second Grand-Challenge and Workshop on Human Multimodal Language**	[Web]

14:20–16:30 **W18: AutoSimTrans (Session 1): The 1st Workshop on Automatic Simultaneous Translation: challenges, recent advances,and future directions** [Web]

15:30–0:00 (+1) **W12: WNGT: The 4th Workshop on Neural Generation and Translation** [Web]

15:30–0:50 (+1) **W11: NLI: Natural Language Interfaces: Challenges and Promises** [Web]

15:30–1:00 (+1) **W14: SIGMORPHON: 17th Workshop on Computational Research in Phonetics, Phonology, and Morphology** [Web]

16:00–1:00 (+1) **W15: NLPMC: NLP for Medical Conversations** [Web]

16:00–19:30 **W2: IWSLT: The 17th International Conference on Spoken Language Translation** [Web]

16:00–23:30 **W17: SocialNLP: The Eighth International Workshop on Natural Language Processing for Social Media** [Web]

22:00–1:30 (+1) **W18: AutoSimTrans (Session 2): The 1st Workshop on Automatic Simultaneous Translation: challenges, recent advances,and future directions** [Web]

WiNLP: The Fourth Widening NLP Workshop

Organizers: *Samira Shaikh, Rossana da Cunha Silva, Ann Clifton, Erika Doggett, and Ryan Georgi*

[Virtual Portal Website] [Workshop Website]

The schedule here may not be up-to-date, please check the workshop website for the latest one.

Sunday, July 5

15:00–16:00 **Sponsored "breakfast"**

16:00–17:30 **Session 1: Opening, Keynote, and Posters**

16:00–16:10 Opening Remarks

16:10–16:50 Keynote 1: Verena Rieser

16:50–17:30 Poster Session A

- Corpus based Amharic sentiment lexicon generation
  Girma Neshir Alemneh, Andreas Rauber, and Solomon Atnafu
- Negation handling for Amharic sentiment classification
  Girma Neshir Alemneh, Andreas Rauber, and Solomon Atnafu
- Embedding Oriented Adaptable Semantic Annotation Framework for Amharic Web Documents
  Kidane Woldemariyam and Dr. Fekade Getahun
- Similarity and Farness Based Bidirectional Neural Co-Attention for Amharic Natural Language Inference
  Abebawu Eshetu, Getenesh Teshome, and Ribka Alemayehu
- Large Vocabulary Read Speech Corpora for Four Ethiopian Languages: Amharic, Tigrigna, Oromo, and Wolaytta
  Solomon Teferra Abate, Martha Yifiru Tachbelie, Michael Melese, Hafte Abera, Tewodros Gebreselassie, Wondwossen Mulugeta, Yaregal Assabie, Million Meshesha Beyene, Solomon Atinafu, and Binyam Ephrem Seyoum
- SIMPLEX-PB 2.0: A Reliable Dataset for Lexical Simplification in Brazilian Portuguese
  Nathan Hartmann, Gustavo Henrique Paetzold, and Sandra Aluísio
- Token Level Identification of Multiword Expressions Using Contextual Information
  REYHANEH HASHEMPOUR and Aline Villavicencio
- Bi-directional Answer-to-Answer Co-attention for Short Answer Grading using Deep Learning
  Abebawu Eshetu, Getenesh Teshome, and Ribka Alemahu
- Effective questions in referential visual dialogue
  Mauricio Mazuecos, Alberto Testoni, Raffaella Bernardi, and Luciana Benotti
- A Translation-Based Approach to Morphology Learning for Low Resource Languages
  Tewodros Gebreselassie, Amanuel Mersha, and Michael Gasser
- Monolingual corpus creation and evaluation of truly low-resource languages from Peru
  Gina Bustamante and Arturo Oncevay
- Tigrinya Automatic Speech recognition with Morpheme based recognition units
  Hafte Abera and sebsibe hailemariam sebsibe
- Variants of Vector Space Reductions for Predicting the Compositionality of English Noun Compounds
  Pegah Alipoormolabashi and Sabine Schulte im Walde
- SentiTel: TABSA for Twitter reviews on Uganda Telecoms
  David Kabiito and Joyce Nakatumba Nabende
- An Assessment of Language Identification Methods on Tweets and Wikipedia Articles
  Pedro Vernetti and Larissa Freitas
- A Comparison of Identification Methods of Brazilian Music Styles by Lyrics
  Patrick Guimarães, Jader Froes, Douglas Costa, and Larissa Freitas
- Enabling fast and correct typing in 'Leichte Sprache' (Easy Language)
  Ina Steinmetz and Karin Harbusch
- AI4D - African Language Dataset Challenge
  Kathleen Siminyu and Sackey Freshia
- Can Wikipedia Categories Improve Masked Language Model Pretraining?
  Diksha Meghwal, Katharina Kann, Iacer Calixto, and Stanislaw Jastrzebski
- Adversarial Evaluation of BERT for Biomedical Named Entity Recognition
  Vladimir Araujo, Andrés Carvallo, and Denis Parra
- FFR v1.1: Fon-French Neural Machine Translation
  Chris Chinenye Emezue and Femi Pancrace Bonaventure Dossou

- Classification and Analysis of Neologisms Produced by Learners of Spanish: Effects of Proficiency and Task
  Shira Wein

17:30–17:50 **Break (sponsored)**

17:50–19:30 **Session 2: Keynote and Mentoring**

17:50–18:30 Keynote 2: Helena Caseli

18:30–19:30 Mentoring and Sponsor Roundtables

19:30–20:10 **Lunch**

20:10–22:10 **Session 3: Panel, Keynote, and Posters**

20:10–20:50 Panel Discussion: Anima Anandkumar, Jade Abbott, Luciana Benotti, and Xanda Schofield

20:50–21:30 Keynote 3: José Eduardo Ochoa

21:30–22:10 Poster Session B

- Developing a Monolingual Sentence Simplification Corpus for Urdu
  Yusra Anees, Sadaf Abdul Rauf, Nauman Iqbal, and Abdul Basit Siddiqi
- Translating Natural Language Instructions for Behavioral Robot Navigation with a Multi-Head Attention Mechanism
  Patricio Cerda-Mardini, Vladimir Araujo, and Álvaro Soto
- Towards Mitigating Gender Bias in a decoder-based Neural Machine Translation model by Adding Contextual Information
  Christine Basta, Marta R. Costa-jussà, and José A. R. Fonollosa
- Predicting and Analyzing Law-Making in Kenya
  Oyinlola Babafemi and Adewale Akinfaderin
- Defining and Evaluating Fair Natural Language Generation
  Catherine Yeo and Alyssa Chen
- Political Advertising Dataset: the use case of the Polish 2020 Presidential Elections
  Lukasz Augustyniak, Krzysztof Rajda, Tomasz Kajdanowicz, and Michał Bernaczyk
- The human unlikeness of neural language models in next-word prediction
  Cassandra L. Jacobs and Arya D. McCarthy
- A Study on the Influence of Architecture Complexity of RNNs for Intent Classification in E-Commerce Chats in Bahasa Indonesia
  Renny Pradina Kusumawardani and Muhammad Azzam
- Long-Tail Predictions with Continuous-Output Language Models
  Shiran Dudy and Steven Bedrick
- Analyzing the Framing of 2020 Presidential Candidates in the News
  Audrey Acken and Dorottya Demszky
- Understanding the Impact of Experiment Design for Evaluating Dialogue System Output
  Sashank Santhanam and Samira Shaikh
- Studying The Effect of Emotional and Moral Language on Information Contagion during the Charlottesville Event
  Khyati Mahajan and Samira Shaikh
- Mapping of Narrative Text Fields To ICD-10 Codes Using Natural Language Processing and Machine Learning
  Risuna Nkolele
- Multitask Models for Controlling the Complexity of Neural Machine Translation
  Sweta Agrawal and Marine Carpuat
- Using Social Media For Bitcoin Day Trading Behavior Prediction
  Anna Paula Pawlicka Maule and Kristen Johnson
- HausaMT v1.0: Towards English—Hausa Neural Machine Translation
  Adewale Akinfaderin
- Outcomes of coming out: Analyzing stories of LGBTQ+
  Krithika Ramesh and Tanvi Anand
- An Evaluation of Subword Segmentation Strategies for Neural Machine Translation of Morphologically Rich Languages
  Aquia Richburg, Ramy Eskander, Smaranda Muresan, and Marine Carpuat
- Enhanced Urdu Word Segmentation using Conditional Random Fields and Morphological Context Features
  Aamir Farhan, Mashrukh Islam, and Dipti Misra Sharma
- Flexible Non-Autoregressive Neural Machine Translation via Repositioning Edit Operations
  Weijia Xu and Marine Carpuat

22:10–22:25 **Break (sponsored)**

22:25–23:15 **Session 4: Keynote and Closing**

22:25–23:05 Keynote 4: Rachael Tatman
23:05–23:15 Closing Remarks

IWSLT: The 17th International Conference on Spoken Language Translation

Organizers: *Marcello Federico, Alexander Waibel, Jiatao Gu, Kevin Knight, Will Lewis, Satoshi Nakamura, Hermann Ney, Jan Niehues, Sebastian Stüker, and Marco Turchi*

[Virtual Portal Website] [Workshop Website]

The schedule here may not be up-to-date, please check the workshop website for the latest one.

Day 1

Evaluation Overview

- FINDINGS OF THE IWSLT 2020 EVALUATION CAMPAIGN
  Ebrahim Ansari, amittai axelrod amittai, Nguyen Bach, Ondřej Bojar, Roldano Cattoni, Fahim Dalvi, Nadir Durrani, Marcello Federico, Christian Federmann, Jiatao Gu, Fei Huang, Kevin Knight, Xutai Ma, Ajay Nagesh, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Xing Shi, Sebastian Stüker, Marco Turchi, Alexander Waibel, and Changhan Wang

Day 1-2

System Papers: Simultaneous Speech Translation

- ON-TRAC Consortium for End-to-End and Simultaneous Speech Translation Challenge Tasks at IWSLT 2020
  Maha Elbayad, Ha Nguyen, Fethi Bougares, Natalia Tomashenko, Antoine Caubrière, Benjamin Lecouteux, Yannick Estève, and Laurent Besacier
- Start-Before-End and End-to-End: Neural Speech Translation by AppTek and RWTH Aachen University
  Parnia Bahar, Patrick Wilken, Tamer Alkhouli, Andreas Guta, Pavel Golik, Evgeny Matusov, and Christian Herold
- KIT's IWSLT 2020 SLT Translation System
  Ngoc-Quan Pham, Felix Schneider, Tuan-Nam Nguyen, Thanh-Le Ha, Thai Son Nguyen, Maximilian Awiszus, Sebastian Stüker, and Alexander Waibel
- End-to-End Simultaneous Translation System for IWSLT2020 Using Modality Agnostic Meta-Learning
  Hou Jeung Han, Mohd Abbas Zaidi, Sathish Reddy Indurthi, Nikhil Kumar Lakumarapu, Beomseok Lee, and Sangha Kim

System Papers: Offline Speech Translation

- DiDi Labs' End-to-end System for the IWSLT 2020 Offline Speech TranslationTask
  Arkady Arkhangorodsky, Yiqi Huang, and amittai axelrod amittai
- End-to-End Offline Speech Translation System for IWSLT 2020 using Modality Agnostic Meta-Learning
  Nikhil Kumar Lakumarapu, Beomseok Lee, Sathish Reddy Indurthi, Hou Jeung Han, Mohd Abbas Zaidi, and Sangha Kim
- End-to-End Speech-Translation with Knowledge Distillation: FBKIWSLT2020
  Marco Gaido, Mattia A. Di Gangi, Matteo Negri, and Marco Turchi
- SRPOL's System for the IWSLT 2020 End-to-End Speech Translation Task
  Tomasz Potapczyk and Pawel Przybysz
- The University of Helsinki Submission to the IWSLT2020 Offline SpeechTranslation Task
  Raúl Vázquez, Mikko Aulamo, Umut Sulubacak, and Jörg Tiedemann
- The AFRL IWSLT 2020 Systems: Work-From-Home Edition
  Brian Ore, Eric Hansen, Tim Anderson, and Jeremy Gwinnup

System Papers: Open Domain Translation

- LIT Team's System Description for Japanese-Chinese Machine Translation Task in IWSLT 2020
  Yimeng Zhuang, Yuan Zhang, and Lijie Wang
- OPPO's Machine Translation System for the IWSLT 2020 Open Domain Translation Task
  Qian Zhang, Xiaopu Li, Dawei Dang, Tingxun Shi, Di Ai, Zhengshan Xue, and Jie Hao
- Character Mapping and Ad-hoc Adaptation: Edinburgh's IWSLT 2020 Open Domain Translation System
  Pinzhen Chen, Nikolay Bogoychev, and Ulrich Germann
- CASIA's System for IWSLT 2020 Open Domain Translation
  Qian Wang, Yuchen Liu, Cong Ma, Yu Lu, Yining Wang, Long Zhou, Yang Zhao, Jiajun Zhang, and Chengqing Zong

- Deep Blue Sonics' Submission to IWSLT 2020 Open Domain Translation Task
  Enmin Su and Yi Ren
- University of Tsukuba's Machine Translation System for IWSLT20 Open Domain Translation Task
  Hongyi Cui, Yizhen Wei, Shohei Iida, Takehito Utsuro, and Masaaki Nagata
- Xiaomi's Submissions for IWSLT 2020 Open Domain Translation Task
  Yuhui Sun, Mengxue Guo, Xiang Li, Jianwei Cui, and Bin Wang
- ISTIC's Neural Machine Translation System for IWSLT'2020
  jiaze wei jiaze, wenbin liu wenbin, zhenfeng wu zhenfeng, you pan you, and yanqing he yanqing
- Octanove Labs' Japanese-Chinese Open Domain Translation System
  Masato Hagiwara

System Papers: Conversational Speech Translation

- NAIST's Machine Translation Systems for IWSLT 2020 Conversational Speech Translation Task
  Ryo Fukuda, Katsuhito Sudoh, and Satoshi Nakamura
- Generating Fluent Translations from Disfluent Text Without Access to Fluent References: IIT BombayIWSLT2020
  Nikhil Saini, Jyotsana Khatri, Preethi Jyothi, and Pushpak Bhattacharyya

System Papers: Video Speech Translation

- The HW-TSC Video Speech Translation System at IWSLT 2020
  Minghan Wang, Hao Yang, Yao Deng, Ying Qin, Lizhi Lei, Daimeng Wei, Hengchao Shang, Ning Xie, Xiaochun Li, and Jiaxian Guo

System Papers: Non-Native Speech Translation

- CUNI Neural ASR with Phoneme-Level Intermediate Step for~Non-Native~SLT at IWSLT 2020
  Peter Polák, Sangeet Sagar, Dominik Macháček, and Ondřej Bojar
- ELITR Non-Native Speech Translation at IWSLT 2020
  Dominik Macháček, Jonáš Kratochvíl, Sangeet Sagar, Matúš Žilinec, Ondřej Bojar, Thai-Son Nguyen, Felix Schneider, Philip Williams, and Yuekun Yao

Research Papers

- Is 42 the Answer to Everything in Subtitling-oriented Speech Translation?
  Alina Karakanta, Matteo Negri, and Marco Turchi
- Re-translation versus Streaming for Simultaneous Translation
  Naveen Arivazhagan, Colin Cherry, Wolfgang Macherey, and George Foster
- Towards Stream Translation: Adaptive Computation Time for Simultaneous Machine Translation
  Felix Schneider and Alexander Waibel
- Neural Simultaneous Speech Translation Using Alignment-Based Chunking
  Patrick Wilken, Tamer Alkhouli, Evgeny Matusov, and Pavel Golik
- Adapting End-to-End Speech Recognition for Readable Subtitles
  Danni Liu, Jan Niehues, and Gerasimos Spanakis
- From Speech-to-Speech Translation to Automatic Dubbing
  Marcello Federico, Robert Enyedi, Roberto Barra-Chicote, Ritwik Giri, Umut Isik, Arvindh Krishnaswamy, and Hassan Sawaf
- Joint Translation and Unit Conversion for End-to-end Localization
  Georgiana Dinu, Prashant Mathur, Marcello Federico, Stanislas Lauly, and Yaser Al-Onaizan
- Efficient Automatic Punctuation Restoration Using Bidirectional Transformers with Robust Inference
  Maury Courtland, Adam Faulkner, and Gayle McElvain
- How Human is Machine Translationese? Comparing Human and Machine Translations of Text and Speech
  Yuri Bizzoni, Tom S Juzek, Cristina España-Bonet, Koel Dutta Chowdhury, Josef van Genabith, and Elke Teich

NLP4ConvAI: NLP for Conversational AI workshop

Organizers: *Tsung-Hsien Wen, Asli Celikyilmaz, Iñigo Casanueva, Mihail Eric, Anuj Kumar, Alexandros Papangelis, Rushin Shah, and Zhou Yu*

[Virtual Portal Website] [Workshop Website]

The schedule here may not be up-to-date, please check the workshop website for the latest one.

Thursday, July 9

13:00 **Opening Remarks**
13:10 Invited Talk (Yun-Nung Chen)
13:40 Invited Talk (Antonie Bordes)
14:10 Invited Talk (Jacob Andreas)
14:40 Using Alternate Representations of Text for Natural Language Understanding
Venkat Varada, Charith Peris, Yangsook Park, and Christopher Dipersio
14:50 On Incorporating Structural Information to improve Dialogue Response Generation
Nikita Moghe, Priyesh Vijayan, Balaraman Ravindran, and Mitesh M. Khapra
15:00 CopyBERT: A Unified Approach to Question Generation with Self-Attention
Stalin Varanasi, Saadullah Amin, and Guenter Neumann
15:10 How to Tame Your Data: Data Augmentation for Dialog State Tracking
Adam Summerville, Jordan Hashemi, James Ryan, and william ferguson william
15:20 Efficient Intent Detection with Dual Sentence Encoders
Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić
15:30 Accelerating Natural Language Understanding in Task-Oriented Dialog
Ojas Ahuja and Shrey Desai
15:40 DLGNet: A Transformer-based Model for Dialogue Response Generation
Olabiyi Oluwatobi and Erik Mueller
15:50 Data Augmentation for Training Dialog Models Robust to Speech Recognition Errors
Longshaokan Wang, Maryam Fazel-Zarandi, Aditya Tiwari, Spyros Matsoukas, and Lazaros Polymenakos

16:00 **Panel**
17:00 Invited Talk (Jesse Thomason)
17:30 Invited Talk (Dilek Hakkani-Tür)
18:00 Invited Talk (Jason Williams)
18:30 Automating Template Creation for Ranking-Based Dialogue Models
Jingxiang Chen, Heba Elfardy, Simi Wang, Andrea Kahn, and Jared Kramer
18:40 From Machine Reading Comprehension to Dialogue State Tracking: Bridging the Gap
Shuyang Gao, Sanchit Agarwal, Di Jin, Tagyoung Chung, and Dilek Hakkani-Tur
18:50 Improving Slot Filling by Utilizing Contextual Information
Amir Pouran Ben Veyseh, Franck Dernoncourt, and Thien Huu Nguyen
19:00 Learning to Classify Intents and Slot Labels Given a Handful of Examples
Jason Krone, Yi Zhang, and Mona Diab
19:10 MultiWOZ 2.2 : A Dialogue Dataset with Additional Annotation Corrections and State Tracking Baselines
Xiaoxue Zang, Abhinav Rastogi, and Jindong Chen
19:20 Sketch-Fill-A-R: A Persona-Grounded Chit-Chat Generation Framework
Michael Shum, Stephan Zheng, Wojciech Kryscinski, Caiming Xiong, and Richard Socher
19:30 Probing Neural Dialog Models for Conversational Understanding
Abdelrhman Saleh, Tovly Deutsch, Stephen Casper, Yonatan Belinkov, and Stuart Shieber

19:40 **Closing Remarks**

BioNLP 2020: Workshop on Biomedical Natural Language Processing

Organizers: *Dina Demner-Fushman, Kevin Cohen, Sophia Ananiadou, and Jun'ichi Tsujii*

[Virtual Portal Website] [Workshop Website]

The schedule here may not be up-to-date, please check the workshop website for the latest one.

Thursday, July 9

15:30–15:40 **Opening remarks**

15:40–17:30 **Session 1: High accuracy information retrieval, spin and bias**

15:40–16:10 Invited Talk – Kirk Roberts

16:10–16:20 Quantifying 60 Years of Gender Bias in Biomedical Research with Word Embeddings
Anthony Rios, Reenam Joshi, and Hejin Shin

16:20–16:30 Sequence-to-Set Semantic Tagging for Complex Query Reformulation and Automated Text Categorization in Biomedical IR using Self-Attention
Manirupa Das, Juanxi Li, Eric Fosler-Lussier, Simon Lin, Steve Rust, Yungui Huang, and Rajiv Ramnath

16:30–16:40 Interactive Extractive Search over Biomedical Corpora
Hillel Taub Tabib, Micah Shlain, Shoval Sadde, Dan Lahav, Matan Eyal, Yaara Cohen, and Yoav Goldberg

16:40–16:50 Improving Biomedical Analogical Retrieval with Embedding of Structural Dependencies
Amandalynne Paullada, Bethany Percha, and Trevor Cohen

16:50–17:00 DeSpin: a prototype system for detecting spin in biomedical publications
Anna Koroleva, Sanjay Kamath, Patrick Bossuyt, and Patrick Paroubek

17:00–17:30 Discussion

17:30–17:45 Coffee Break

17:45–20:00 **Session 2: Clinical Language Processing**

17:45–18:15 Invited Talk – Tim Miller

18:15–18:25 Towards Visual Dialog for Radiology
Olga Kovaleva, Chaitanya Shivade, Satyananda Kashyap, Karina Kanjaria, Joy Wu, Deddeh Ballah, Adam Coy, Alexandros Karargyris, Yufan Guo, David Beymer Beymer, Anna Rumshisky, and Vandana Mukherjee Mukherjee

18:25–18:35 A BERT-based One-Pass Multi-Task Model for Clinical Temporal Relation Extraction
Chen Lin, Timothy Miller, Dmitriy Dligach, Farig Sadeque, Steven Bethard, and Guergana Savova

18:35–18:45 Experimental Evaluation and Development of a Silver-Standard for the MIMIC-III Clinical Coding Dataset
Thomas Searle, Zina Ibrahim, and Richard Dobson

18:45–18:55 Comparative Analysis of Text Classification Approaches in Electronic Health Records
Aurelie Mascio, Zeljko Kraljevic, Daniel Bean, Richard Dobson, Robert Stewart, Rebecca Bendayan, and Angus Roberts

18:55–19:05 Noise Pollution in Hospital Readmission Prediction: Long Document Classification with Reinforcement Learning
Liyan Xu, Julien Hogan, Rachel E. Patzer, and Jinho D. Choi

19:05–19:15 Evaluating the Utility of Model Configurations and Data Augmentation on Clinical Semantic Textual Similarity
Yuxia Wang, Fei Liu, Karin Verspoor, and Timothy Baldwin

19:15–19:45 Discussion

19:45–20:30 Lunch

20:30–22:30 **Session 3: Language Understanding**

20:30–21:00 Invited Talk – Anna Rumshisky

21:00–21:10 Entity-Enriched Neural Models for Clinical Question Answering
Bhanu Pratap Singh Rawat, Wei-Hung Weng, So Yeon Min, Preethi Raghavan, and Peter Szolovits

21:10–21:20 Evidence Inference 2.0: More Data, Better Models
Jay DeYoung, Eric Lehman, Benjamin Nye, Iain Marshall, and Byron C. Wallace

21:20–21:30 Personalized Early Stage Alzheimer's Disease Detection: A Case Study of President Reagan's Speeches
Ning Wang, Fan Luo, Vishal Peddagangireddy, Koduvayur Subbalakshmi, and Rajarathnam Chandramouli

21:30–21:40 BioMRC: A Dataset for Biomedical Machine Reading Comprehension
Dimitris Pappas, Petros Stavropoulos, Ion Androutsopoulos, and Ryan McDonald

21:40–21:50 Neural Transduction of Letter Position Dyslexia using an Anagram Matrix Representation
Avi Bleiweiss

21:50–22:00 Domain Adaptation and Instance Selection for Disease Syndrome Classification over Veterinary Clinical Notes
Brian Hur, Timothy Baldwin, Karin Verspoor, Laura Hardefeldt, and James Gilkerson

22:00–22:30 Discussion

22:30–22:45 Coffee Break

22:45–0:45 (+1) **Session 4: Named Entity Recognition and Knowledge Representation**

22:45–23:25 Invited Talk: Machine Reading for Precision Medicine, Hoifung Poon

23:25–23:35 Benchmark and Best Practices for Biomedical Knowledge Graph Embeddings
David Chang, Ivana Balažević, Carl Allen, Daniel Chawla, Cynthia Brandt, and Andrew Taylor

23:35–23:45 Extensive Error Analysis and a Learning-Based Evaluation of Medical Entity Recognition Systems to Approximate User Experience
Isar Nejadgholi, Kathleen C. Fraser, and Berry de Bruijn

23:45–23:55 A Data-driven Approach for Noise Reduction in Distantly Supervised Biomedical Relation Extraction
Saadullah Amin, Katherine Ann Dunfield, Anna Vechkaeva, and Guenter Neumann

23:55–0:05 (+1) Global Locality in Biomedical Relation and Event Extraction
Elaheh ShafieiBavani, Antonio Jimeno Yepes, Xu Zhong, and David Martinez Iraola

Friday, July 10

0:05–0:15 An Empirical Study of Multi-Task Learning on BERT for Biomedical Text Mining
Yifan Peng, Qingyu Chen, and Zhiyong Lu

0:15–0:45 Discussion

0:45–1:00 **Closing remarks**

FEVER: The Third workshop on Fact Extraction and VERification

Organizers: *Christos Christodoulopoulos, James Thorne, Andreas Vlachos, Oana Cocarascu, and Arpit Mittal*

[Virtual Portal Website] [Workshop Website]

The schedule here may not be up-to-date, please check the workshop website for the latest one.

Opening Remarks (FEVER Organizers)

Project Debater (Noam Slonim)

Towards explainable fact checking (Isabelle Augenstein)

Oral presentations

- Simple Compounded-Label Training for Fact Extraction and Verification
  Yixin Nie, Lisa Bauer, and Mohit Bansal
- Stance Prediction and Claim Verification: An Arabic Perspective
  Jude Khouja

How to "inoculate" people against misinformation and online extremism (Jon Roozenbeek)

Beyond Facts: The Problem of Framing in Assessing What is True (Phil Resnik)

Poster Session

- A Probabilistic Model with Commonsense Constraints for Pattern-based Temporal Fact Extraction
  Yang Zhou, Tong Zhao, and Meng Jiang
- Developing a How-to Tip Machine Comprehension Dataset and its Evaluation in Machine Comprehension by BERT
  Tengyang Chen, Hongyu Li, Miho Kasamatsu, Takehito Utsuro, and Yasuhide Kawada
- Language Models as Fact Checkers?
  Nayeon Lee, Belinda Li, Sinong Wang, Wen-tau Yih, Hao Ma, and Madian Khabsa
- Maintaining Quality in FEVER Annotation
  Leon Derczynski, Julie Binau, and Henri Schulte
- Distilling the Evidence to Augment Fact Verification Models
  Beatrice Portelli, Jason Zhao, Tal Schuster, Giuseppe Serra, and Enrico Santus

Integration of (Un)structured World Knowledge In Task Oriented Conversations (Dilek Hakkani-Tur)

Closing Remarks (FEVER Organizers)

IWPT: The 16th International Conference on Parsing Technologies

Organizers: *Yuji Matsumoto, Stephan Oepen, Kenji Sagae, Anders Søgaard, Weiwei Sun, and Reut Tsarfaty*

[Virtual Portal Website] [Workshop Website]

The schedule here may not be up-to-date, please check the workshop website for the latest one.

Thursday, July 9

14:00–14:15 **Session 1: Invited Talk Q&A**

- Syntactic Parsing in Humans and Machines
  Paola Merlo

14:15–14:40 **Session 2: Regular Papers Q&A**

- Distilling Neural Networks for Greener and Faster Dependency Parsing
  Mark Anderson and Carlos Gómez-Rodríguez
- End-to-End Negation Resolution as Graph Parsing
  Robin Kurtz, Stephan Oepen, and Marco Kuhlmann
- Integrating Graph-Based and Transition-Based Dependency Parsers in the Deep Contextualized Era
  Agnieszka Falenska, Anders Björkelund, and Jonas Kuhn
- Semi-supervised Parsing with a Variational Autoencoding Parser
  Xiao Zhang and Dan Goldwasser

14:40–15:00 **Session 3: Regular Papers Q&A**

- Memory-bounded Neural Incremental Parsing for Psycholinguistic Prediction
  Lifeng Jin and William Schuler
- Obfuscation for Privacy-preserving Syntactic Parsing
  Zhifeng Hu, Serhii Havrylov, Ivan Titov, and Shay B. Cohen
- Tensors over Semirings for Latent-Variable Weighted Logic Programs
  Esma Balkir, Daniel Gildea, and Shay B. Cohen

15:10–15:35 **Session 4: Regular Papers Q&A**

- Advances in Using Grammars with Latent Annotations for Discontinuous Parsing
  Kilian Gebhardt
- Lexicalization of Probabilistic Linear Context-free Rewriting Systems
  Richard Mörbitz and Thomas Ruprecht
- Self-Training for Unsupervised Parsing with PRPN
  Anhad Mohananey, Katharina Kann, and Samuel R. Bowman
- Span-Based LCFRS-2 Parsing
  Miloš Stanojević and Mark Steedman

15:35–16:00 **Session 5: Regular Papers Q&A**

- Analysis of the Penn Korean Universal Dependency Treebank (PKT-UD): Manual Revision to Build Robust Parsing Model in Korean
  Tae Hwan Oh, Ji Yoon Han, Hyonsu Choe, Seokwon Park, Han He, Jinho D. Choi, Na-Rae Han, Jena D. Hwang, and Hansaem Kim
- Statistical Deep Parsing for Spanish Using Neural Networks
  Luis Chiruzzo and Dina Wonsever
- The Importance of Category Labels in Grammar Induction with Child-directed Utterances
  Lifeng Jin and William Schuler

16:10–17:00 **Session 6: Shared Task Q&A**

- Overview of the IWPT 2020 Shared Task on Parsing into Enhanced Universal Dependencies
  Gosse Bouma, Djamé Seddah, and Daniel Zeman
- Turku Enhanced Parser Pipeline: From Raw Text to Enhanced Graphs in the IWPT 2020 Shared Task
  Jenna Kanerva, Filip Ginter, and Sampo Pyysalo
- Hybrid Enhanced Universal Dependencies Parsing
  Johannes Heinecke
- Adaptation of Multilingual Transformer Encoder for Robust Enhanced Universal Dependency Parsing
  Han He and Jinho D. Choi

- Efficient EUD Parsing
  Mathieu Dehouck, Mark Anderson, and Carlos Gómez-Rodríguez
- Linear Neural Parsing and Hybrid Enhancement for Enhanced Universal Dependencies
  Giuseppe Attardi, Daniele Sartiano, and Maria Simi
- Enhanced Universal Dependency Parsing with Second-Order Inference and Mixture of Training Data
  Xinyu Wang, Yong Jiang, and Kewei Tu
- How Much of Enhanced UD Is Contained in UD?
  Adam Ek and Jean-Philippe Bernardy
- The ADAPT Enhanced Dependency Parser at the IWPT 2020 Shared Task
  James Barry, Joachim Wagner, and Jennifer Foster
- Køpsala: Transition-Based Graph Parsing via Efficient Training and Effective Encoding
  Daniel Hershcovich, Miryam de Lhoneux, Artur Kulmizev, Elham Pejhan, and Joakim Nivre
- RobertNLP at the IWPT 2020 Shared Task: Surprisingly Simple Enhanced UD Parsing for English
  Stefan Grünewald and Annemarie Friedrich

FLP: The 2nd Workshop on Figurative Language Processing

Organizers: *Beata Beigman Klebanov, Ekaterina Shutova, Patricia Lichtenstein, Smaranda Muresan, Anna Feldman, Chee Wee (Ben) Leong, and Debanjan Ghosh*

[Virtual Portal Website] [Workshop Website]

The schedule here may not be up-to-date, please check the workshop website for the latest one.

Thursday, July 9

13:05 A Report on the 2020 Sarcasm Detection Shared Task
Debanjan Ghosh, Avijit Vajpayee, and Smaranda Muresan

13:20 Augmenting Data for Sarcasm Detection with Unlabeled Conversation Context
Hankyol Lee, Youngjae Yu, and Gunhee Kim

13:35 A Report on the 2020 VUA and TOEFL Metaphor Detection Shared Task
Chee Wee (Ben) Leong, Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and Xianyang Chen

13:50 DeepMet: A Reading Comprehension Paradigm for Token-level Metaphor Detection
Chuandong Su, Fumiyo Fukumoto, Xiaoxi Huang, Jiyi Li, Rongbo Wang, and Zhiqun Chen

14:05 Context-Driven Satirical News Generation
Zachary Horvitz, Nam Do, and Michael L. Littman

14:40 Sarcasm Detection using Context Separators in Online Discourse
TANVI DADU and Kartikey Pant

14:45 Sarcasm Detection in Tweets with BERT and GloVe Embeddings
Akshay Khatri and Pranav P

14:50 C-Net: Contextual Network for Sarcasm Detection
Amit Kumar Jena, Aman Sinha, and Rohit Agarwal

15:00 Applying Transformers and Aspect-based Sentiment Analysis approaches on Sarcasm Detection
Taha Shangipour ataei, Soroush Javdan, and Behrouz Minaei-Bidgoli

15:05 Sarcasm Identification and Detection in Conversion Context using BERT
kalaivani A and Thenmozhi D

15:10 Neural Sarcasm Detection using Conversation Context
Nikhil Jaiswal

15:40 Context-Aware Sarcasm Detection Using BERT
Arup Baruah, Kaushik Das, Ferdous Barbhuiya, and Kuntal Dey

15:45 Transformers on Sarcasm Detection with Context
Amardeep Kumar and Vivek Anand

15:50 A Novel Hierarchical BERT Architecture for Sarcasm Detection
Himani Srivastava, Vaibhav Varshney, Surabhi Kumari, and Saurabh Srivastava

16:00 Detecting Sarcasm in Conversation Context Using Transformer-Based Models
Adithya Avvaru, Sanath Vobilisetty, and Radhika Mamidi

16:05 Using Conceptual Norms for Metaphor Detection
Mingyu WAN, Kathleen Ahrens, Emmanuele Chersoni, Menghan Jiang, Qi Su, Rong Xiang, and Chu-Ren Huang

16:10 ALBERT-BiLSTM for Sequential Metaphor Detection
Shuqun Li, Jingjie Zeng, Jinhui Zhang, Tao Peng, Liang Yang, and Hongfei Lin

16:15 Character aware models with similarity learning for metaphor detection
Tarun Kumar and Yashvardhan Sharma

17:00 Sky + Fire = Sunset. Exploring Parallels between Visually Grounded Metaphors and Image Classifiers
Yuri Bizzoni and Simon Dobnik

17:15 Recognizing Euphemisms and Dysphemisms Using Sentiment Analysis
Christian Felt and Ellen Riloff

17:30 IlliniMet: Illinois System for Metaphor Detection with Contextual and Linguistic Information
Hongyu Gong, Kshitij Gupta, Akriti Jain, and Suma Bhat

17:35 Adaptation of Word-Level Benchmark Datasets for Relation-Level Metaphor Identification
Omnia Zayed, John Philip McCrae, and Paul Buitelaar

17:40 Generating Ethnographic Models from Communities' Online Data
Tomek Strzalkowski, Anna Newheiser, Nathan Kemper, Ning Sa, Bharvee Acharya, and Gregorios Katsios

17:50 Oxymorons: a preliminary corpus investigation
Marta La Pietra and Francesca Masini

17:55 Can Humor Prediction Datasets be used for Humor Generation? Humorous Headline Generation via Style Transfer
Orion Weller, Nancy Fulda, and Kevin Seppi

18:00 Evaluating a Bi-LSTM Model for Metaphor Detection in TOEFL Essays
Kevin Kuo and Marine Carpuat

18:30 Neural Metaphor Detection with a Residual biLSTM-CRF Model
Andrés Torres Rivera, Antoni Oliver, Salvador Climent, and Marta Coll-Florit

18:35 Augmenting Neural Metaphor Detection with Concreteness
Ghadi Alnafesah, Harish Tayyar Madabushi, and Mark Lee

18:40 Supervised Disambiguation of German Verbal Idioms with a BiLSTM Architecture
Rafael Ehren, Timm Lichte, Laura Kallmeyer, and Jakub Waszczuk

18:50 Metaphor Detection using Context and Concreteness
Rowan Hall Maudslay, Tiago Pimentel, Ryan Cotterell, and Simone Teufel

18:55 Being neighbourly: Neural metaphor identification in discourse
Verna Dankers, Karan Malhotra, Gaurav Kudva, Volodymyr Medentsiy, and Ekaterina Shutova

19:00 Go Figure! Multi-task transformer-based architecture for metaphor detection using idioms: ETS team in 2020 metaphor shared task
Xianyang Chen, Chee Wee (Ben) Leong, Michael Flor, and Beata Beigman Klebanov

19:10 Metaphor Detection using Ensembles of Bidirectional Recurrent Neural Networks
Jennifer Brooks and Abdou Youssef

19:15 Metaphor Detection Using Contextual Word Embeddings From Transformers
Jerry Liu, Nathan O'Hara, Alexander Rubin, Rachel Draelos, and Cynthia Rudin

19:20 Testing the role of metadata in metaphor identification
Egon Stemle and Alexander Onysko

19:50 Sarcasm Detection Using an Ensemble Approach
Jens Lemmens, Ben Burtenshaw, Ehsan Lotfi, Ilia Markov, and Walter Daelemans

19:55 A Transformer Approach to Contextual Sarcasm Detection in Twitter
Hunter Gregory, Steven Li, Pouya Mohammadi, Natalie Tarn, Rachel Draelos, and Cynthia Rudin

20:00 Transformer-based Context-aware Sarcasm Detection in Conversation Threads from Social Media
Xiangjue Dong, Changmao Li, and Jinho D. Choi

NUSE: The 1st Joint Workshop on Narrative Understanding, Storylines, and Events

Organizers: *Claire Bonial, Tommaso Caselli, Snigdha Chaturvedi, Elizabeth Clark, Ruihong Huang, Ben Miller, Mohit Iyyer, Alejandro Jaimes, Heng Ji, Lara Martin, Teruko Mitamura, Nanyun Peng, and Joel Tetreault*

[Virtual Portal Website] [Workshop Website]

The schedule here may not be up-to-date, please check the workshop website for the latest one.

Thursday, July 9

15:15–15:30 **Opening Remarks**

15:30–16:30 **Q+A (Session 1)**

17:00–18:00 **Speaker Panel (Session 1)**

19:00–20:30 **Annotation Exercise**

22:00–23:00 **Speaker Panel (Session 2)**

23:00–0:00 (+1) **Q+A (Session 2)**

Friday, July 10

0:00–0:15 **Closing Remarks**

Keynote Talks

Keynote 1 (Angela Fan)
Keynote 2 (Mark Finlayson)
Keynote 3 (Andrew Gordon)
Keynote 4 (Alexander G. Hauptmann)
Keynote 5 (Kathleen McKeown)
Keynote 6 (Ellen Riloff)
Keynote 7 (Ted Underwood)

Paper Talks (Archival)

- New Insights into Cross-Document Event Coreference: Systematic Comparison and a Simplified Approach
  Andres Cremisini and Mark Finlayson
- Screenplay Quality Assessment: Can We Predict Who Gets Nominated?
  Ming-Chang Chiu, Tiantian Feng, Xiang Ren, and Shrikanth Narayanan
- Improving the Identification of the Discourse Function of News Article Paragraphs
  Deya Banisakher, W. Victor Yarlott, Mohammed Aldawsari, Naphtali Rishe, and Mark Finlayson
- Systematic Evaluation of a Framework for Unsupervised Emotion Recognition for Narrative Text
  Samira Zad and Mark Finlayson
- Extensively Matching for Few-shot Learning Event Detection
  Viet Dac Lai, Thien Huu Nguyen, and Frank Dernoncourt
- Exploring the Effect of Author and Reader Identity in Online Story Writing: the STORIESINTHEWILD Corpus.
  Tal August, Maarten Sap, Elizabeth Clark, Katharina Reinecke, and Noah A. Smith
- Script Induction as Association Rule Mining
  Anton Belyy and Benjamin Van Durme
- Automatic extraction of personal events from dialogue
  Joshua Eisenberg and Michael Sheriff
- Annotating and quantifying narrative time disruptions in modernist and hypertext fiction
  Edward Kearns
- Exploring aspects of similarity between spoken personal narratives by disentangling them into narrative clause types
  Belen Saldias and Deb Roy
- Extracting Message Sequence Charts from Hindi Narrative Text
  Swapnil Hingmire, Nitin Ramrakhiyani, Avinash Kumar Singh, Sangameshwar Patil, Girish Palshikar, Pushpak Bhattacharyya, and Vasudeva Varma
- Emotion Arcs of Student Narratives
  Swapna Somasundaran, Xianyang Chen, and Michael Flor

- Frustratingly Hard Evidence Retrieval for QA Over Books
  Xiangyang Mou, Mo Yu, Bingsheng Yao, Chenghao Yang, Xiaoxiao Guo, Saloni Potdar, and Hui Su
- On-The-Fly Information Retrieval Augmentation for Language Models
  Hai Wang and David McAllester
- Detecting and understanding moral biases in news
  Usman Shahid, Barbara Di Eugenio, Andrew Rojecki, and Elena Zheleva

Paper Talks (Non-Archival)

Bringing Stories Alive: Generating Interactive Fiction Worlds (Prithviraj Ammanabrolu, Wesley Cheung, Dan Tu, William Broniec and Mark Riedl)

CompRes: A Dataset for Narrative Structure in News (Effi Levi, Guy Mor, Shaul Shenhav and Tamir Sheafer)

ALVR: Workshop on Advances in Language and Vision Research

Organizers: *Xin Wang, Jesse Thomason, Ronghang Hu, Xinlei Chen, Peter Anderson, Qi Wu, Asli Celikyilmaz, Jason Baldridge, and William Yang Wang*

[Virtual Portal Website] [Workshop Website]

The schedule here may not be up-to-date, please check the workshop website for the latest one.

- Extending ImageNet to Arabic using Arabic WordNet
  Abdulkareem Alsudais
- Toward General Scene Graph: Integration of Visual Semantic Knowledge with Entity Synset Alignment
  Woo Suk Choi, Kyoung-Woon On, Yu-Jung Heo, and Byoung-Tak Zhang
- Visual Question Generation from Radiology Images
  Mourad Sarrouti, Asma Ben Abacha, and Dina Demner-Fushman
- On the role of effective and referring questions in GuessWhat?!
  Mauricio Mazuecos, Alberto Testoni, Raffaella Bernardi, and Luciana Benotti
- Latent Alignment of Procedural Concepts in Multimodal Recipes
  Hossein Rajaby Faghihi, Roshanak Mirzaee, Sudarshan Paliwal, and Parisa Kordjamshidi

RepL4NLP: The 5th Workshop on Representation Learning for NLP

Organizers: *Emma Strubell, Spandana Gella, Marek Rei, Johannes Welbl, Fabio Petroni, Patrick Lewis, Hannaneh Hajishirzi, Kyunghyun Cho, Edward Grefenstette, Karl Moritz Hermann, Laura Rimell, Chris Dyer, and Isabelle Augenstein*

[Virtual Portal Website] [Workshop Website]

The schedule here may not be up-to-date, please check the workshop website for the latest one.

Thursday, July 9

8:00–9:45 **Session 1**

8:00–8:15 Welcome and Opening Remarks

8:15–9:45 Poster Session 1

- Improving Bilingual Lexicon Induction with Unsupervised Post-Processing of Monolingual Word Vector Spaces
  Ivan Vulić, Anna Korhonen, and Goran Glavaš

  On the Ability of Self-Attention Networks to Recognize Counter Languages (Satwik Bhattamishra, Kabir Ahuja and Navin Goyal)
- Word Embeddings as Tuples of Feature Probabilities
  Siddharth Bhat, Alok Debnath, Souvik Banerjee, and Manish Shrivastava
- Compositionality and Capacity in Emergent Languages
  Abhinav Gupta, Cinjon Resnick, Jakob Foerster, Andrew Dai, and Kyunghyun Cho
- Learning Geometric Word Meta-Embeddings
  Pratik Jawanpuria, Satya Dev N T V, Anoop Kunchukuttan, and Bamdev Mishra

  Variational Inference for Learning Representations of Natural Language Edits (Edison Marrese-Taylor, Machel Reid and Yutaka Matsuo)
- Adversarial Training for Commonsense Inference
  Lis Pereira, Xiaodong Liu, Fei Cheng, Masayuki Asahara, and Ichiro Kobayashi
- Joint Training with Semantic Role Labeling for Better Generalization in Natural Language Inference
  Cemil Cengiz and Deniz Yuret
- A Metric Learning Approach to Misogyny Categorization
  Juan Manuel Coria, Sahar Ghannay, Sophie Rosset, and Hervé Bredin
- On the Choice of Auxiliary Languages for Improved Sequence Tagging
  Lukas Lange, Heike Adel, and Jannik Strötgen
- Adversarial Alignment of Multilingual Models for Extracting Temporal Expressions from Text
  Lukas Lange, Anastasiia Iurshina, Heike Adel, and Jannik Strötgen
- Contextual and Non-Contextual Word Embeddings: an in-depth Linguistic Investigation
  Alessio Miaschi and Felice Dell'Orletta
- Staying True to Your Word: (How) Can Attention Become Explanation?
  Martin Tutek and Jan Snajder

  A Simple Approach to Learning Unsupervised Multilingual Embeddings (Pratik Jawanpuria, Mayank Meghwanshi and Bamdev)
- What's in a Name? Are BERT Named Entity Representations just as Good for any other Name?
  Sriram Balasubramanian, Naman Jain, Gaurav Jindal, Abhijeet Awasthi, and Sunita Sarawagi

  AI4Bharat-IndicNLP Dataset: Monolingual Corpora and Word Embeddings for Indic Languages: Monolingual Corpora and Word Embeddings for Indic Languages (Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra and Pratyush Kumar)
- Evaluating Natural Alpha Embeddings on Intrinsic and Extrinsic Tasks
  Riccardo Volpi and Luigi Malagò

  Predicting Sexual and Reproductive Health of Migrants using Data Science (Amber Nigam, Pragati Jaiswal, Teertha Arora, Uma Girkar and Leo Anthony Celi)

  Job Recommendation through Progression of Job Selection (Amber Nigam, Aakash Roy, Hartaran Singh and Arpan Saxena)

14:00–17:00 **Session 2**

14:00–14:15 Welcome and Opening Remarks

14:15–14:45 Invited speaker Q&A: I don't know what you mean semantics is hard: Challenges in evaluation of semantic phenomena (Ellie Pavlick)

14:45–15:00 Break

15:00–15:30 Invited speaker Q&A: Artificial Neural Networks as models of language comprehension in the human brain (Evelina Fedorenko)

15:30–17:00 Poster Session 2

- Improving Bilingual Lexicon Induction with Unsupervised Post-Processing of Monolingual Word Vector Spaces
  Ivan Vulić, Anna Korhonen, and Goran Glavaš
- Are All Languages Created Equal in Multilingual BERT?
  Shijie Wu and Mark Dredze

  On the Ability of Self-Attention Networks to Recognize Counter Languages (Satwik Bhattamishra, Kabir Ahuja and Navin Goyal)
- Zero-Resource Cross-Domain Named Entity Recognition
  Zihan Liu, Genta Indra Winata, and Pascale Fung
- Encodings of Source Syntax: Similarities in NMT Representations Across Target Languages
  Tyler A. Chang and Anna Rafferty
- Learning Probabilistic Sentence Representations from Paraphrases
  Mingda Chen and Kevin Gimpel
- Word Embeddings as Tuples of Feature Probabilities
  Siddharth Bhat, Alok Debnath, Souvik Banerjee, and Manish Shrivastava
- Compositionality and Capacity in Emergent Languages
  Abhinav Gupta, Cinjon Resnick, Jakob Foerster, Andrew Dai, and Kyunghyun Cho
- Learning Geometric Word Meta-Embeddings
  Pratik Jawanpuria, Satya Dev N T V, Anoop Kunchukuttan, and Bamdev Mishra
- Joint Training with Semantic Role Labeling for Better Generalization in Natural Language Inference
  Cemil Cengiz and Deniz Yuret
- A Metric Learning Approach to Misogyny Categorization
  Juan Manuel Coria, Sahar Ghannay, Sophie Rosset, and Hervé Bredin
- On the Choice of Auxiliary Languages for Improved Sequence Tagging
  Lukas Lange, Heike Adel, and Jannik Strötgen
- Adversarial Alignment of Multilingual Models for Extracting Temporal Expressions from Text
  Lukas Lange, Anastasiia Iurshina, Heike Adel, and Jannik Strötgen
- Contextual and Non-Contextual Word Embeddings: an in-depth Linguistic Investigation
  Alessio Miaschi and Felice Dell'Orletta
- Staying True to Your Word: (How) Can Attention Become Explanation?
  Martin Tutek and Jan Snajder
- Compressing BERT: Studying the Effects of Weight Pruning on Transfer Learning
  Mitchell Gordon, Kevin Duh, and Nicholas Andrews
- On Dimensional Linguistic Properties of the Word Embedding Space
  Vikas Raunak, Vaibhav Kumar, Vivek Gupta, and Florian Metze
- A Cross-Task Analysis of Text Span Representations
  Shubham Toshniwal, Haoyue Shi, Bowen Shi, Lingyu Gao, Karen Livescu, and Kevin Gimpel
- Enhancing Transformer with Sememe Knowledge
  Yuhui Zhang, Chenghao Yang, Zhengping Zhou, and Zhiyuan Liu
- Evaluating Compositionality of Sentence Representation Models
  Hanoz Bhathena, Angelica Willis, and Nathan Dass
- Supertagging with CCG primitives
  Aditya Bhargava and Gerald Penn
- What's in a Name? Are BERT Named Entity Representations just as Good for any other Name?
  Sriram Balasubramanian, Naman Jain, Gaurav Jindal, Abhijeet Awasthi, and Sunita Sarawagi
- Exploring the Limits of Simple Learners in Knowledge Distillation for Document Classification with DocBERT
  Ashutosh Adhikari, Achyudh Ram, Raphael Tang, William L. Hamilton, and Jimmy Lin

  AI4Bharat-IndicNLP Dataset: Monolingual Corpora and Word Embeddings for Indic Languages: Monolingual Corpora and Word Embeddings for Indic Languages (Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra and Pratyush Kumar)
- Evaluating Natural Alpha Embeddings on Intrinsic and Extrinsic Tasks
  Riccardo Volpi and Luigi Malagò

  Predicting Sexual and Reproductive Health of Migrants using Data Science (Amber Nigam, Pragati Jaiswal, Teertha Arora, Uma Girkar and Leo Anthony Celi)

  Job Recommendation through Progression of Job Selection (Amber Nigam, Aakash Roy, Hartaran Singh and Arpan Saxena)

Friday, July 10

0:00–3:00 **Session 3**

0:00–0:15 Welcome and Opening Remarks

0:15–0:45 Invited speaker Q&A: Text representations for retrieval and question answering (Kristina Toutanova)

0:45–1:00 Break

1:00–1:30 Invited speaker Q&A: Beyond BERT (Mike Lewis)

1:30–3:00 Poster Session 3

- Are All Languages Created Equal in Multilingual BERT?
  Shijie Wu and Mark Dredze
- Zero-Resource Cross-Domain Named Entity Recognition
  Zihan Liu, Genta Indra Winata, and Pascale Fung
- Encodings of Source Syntax: Similarities in NMT Representations Across Target Languages
  Tyler A. Chang and Anna Rafferty
- Learning Probabilistic Sentence Representations from Paraphrases
  Mingda Chen and Kevin Gimpel

  Variational Inference for Learning Representations of Natural Language Edits (Edison Marrese-Taylor, Machel Reid and Yutaka Matsuo)
- Adversarial Training for Commonsense Inference
  Lis Pereira, Xiaodong Liu, Fei Cheng, Masayuki Asahara, and Ichiro Kobayashi
- Compressing BERT: Studying the Effects of Weight Pruning on Transfer Learning
  Mitchell Gordon, Kevin Duh, and Nicholas Andrews
- On Dimensional Linguistic Properties of the Word Embedding Space
  Vikas Raunak, Vaibhav Kumar, Vivek Gupta, and Florian Metze
- A Cross-Task Analysis of Text Span Representations
  Shubham Toshniwal, Haoyue Shi, Bowen Shi, Lingyu Gao, Karen Livescu, and Kevin Gimpel
- Enhancing Transformer with Sememe Knowledge
  Yuhui Zhang, Chenghao Yang, Zhengping Zhou, and Zhiyuan Liu
- Evaluating Compositionality of Sentence Representation Models
  Hanoz Bhathena, Angelica Willis, and Nathan Dass
- Supertagging with CCG primitives
  Aditya Bhargava and Gerald Penn
- Exploring the Limits of Simple Learners in Knowledge Distillation for Document Classification with DocBERT
  Ashutosh Adhikari, Achyudh Ram, Raphael Tang, William L. Hamilton, and Jimmy Lin

NLI: Natural Language Interfaces: Challenges and Promises

Organizers: *Ahmed Hassan Awadallah, Yu Su, Huan Sun, and Scott Wen-tau Yih*

[Virtual Portal Website] [Workshop Website]

The schedule here may not be up-to-date, please check the workshop website for the latest one.

Friday, July 10

15:15–15:30 Opening remarks

15:30–16:30 Invited Talk: Joyce Chai

16:30–17:30 Invited Talk: Imed Zitouni: On the Path to Billion Skills: Turning Documents into Actions

17:30–17:45 Break

17:45–18:00 Answering Complex Questions by Combining Information from Curated and Extracted Knowledge Bases
Nikita Bhutani, Xinyi Zheng, Kun Qian, Yunyao Li, and H. Jagadish

18:00–18:15 Unnatural Language Processing: Bridging the Gap Between Synthetic and Natural Language Data (Cross-submission) (Alana Marzoev, Samuel Madden, M. Frans Kaashoek, Michael Cafarella and Jacob Andreas)

18:15–18:30 Towards Reversal-Based Textual Data Augmentation for NLI Problems with Opposable Classes
Alexey Tarasov

18:30–19:30 Invited Talk: Monica Lam

19:30–20:30 Break

20:30–21:30 Invited Talk: Percy Liang: Reflections on Semantic Parsing and Learning from Users

21:30–21:45 Examination and Extension of Strategies for Improving Personalized Language Modeling via Interpolation
Liqun Shao, Sahitya Mantravadi, Tom Manzini, Alejandro Buendia, Manon Knoertzer, Soundar Srinivasan, and Chris Quirk

21:45–22:00 A Multi-Modal Agent that Learns Concepts and Conditionals from Natural Language and Demonstrations (Cross-submussion) (Toby Jia-Jun Li, Marissa Radensky, Justin Jia, Kirielle Singarajah, Tom Mitchell and Brad Myers)

22:00–22:15 Efficient Deployment of Conversational Natural Language Interfaces over Databases
Anthony Colas, Trung Bui, Franck Dernoncourt, Moumita Sinha, and Doo Soon Kim

22:15–22:30 Neural Multi-task Text Normalization and Sanitization with Pointer-Generator
Hoang Nguyen and Sandro Cavallari

22:30–22:45 Break

22:45–23:45 Invited Talk: H V Jagadish: Natural Language In A Database Management System

23:45–0:45 (+1) Invited Talk: Luke Zettlemoyer: Denoising Sequence-to-Sequence Pre-training

Saturday, July 11

0:45–0:50 Closing remarks

WNGT: The 4th Workshop on Neural Generation and Translation

Organizers: *Alexandra Birch, Graham Neubig, Andrew Finch, Hiroaki Hayashi, Kenneth Heafield, Ioannis Konstas, Yusuke Oda, and Xian Li*

[Virtual Portal Website] [Workshop Website]

The schedule here may not be up-to-date, please check the workshop website for the latest one.

- Findings of the Fourth Workshop on Neural Generation and Translation
  Kenneth Heafield, Hiroaki Hayashi, Yusuke Oda, Ioannis Konstas, Andrew Finch, Graham Neubig, Xian Li, and Alexandra Birch
- Learning to Generate Multiple Style Transfer Outputs for an Input Sentence
  Kevin Lin, Ming-Yu Liu, Ming-Ting Sun, and Jan Kautz
- Balancing Cost and Benefit with Tied-Multi Transformers
  Raj Dabre, Raphael Rubino, and Atsushi Fujita
- Compressing Neural Machine Translation Models with 4-bit Precision
  Alham Fikri Aji and Kenneth Heafield
- Meta-Learning for Few-Shot NMT Adaptation
  Amr Sharaf, Hany Hassan, and Hal Daumé III
- Automatically Ranked Russian Paraphrase Corpus for Text Generation
  Vadim Gudkov, Olga Mitrofanova, and Elizaveta Filippskikh
- Increasing Lexical Diversity in Plug and Play Language Models
  Soham Parikh, Daphne Ippolito, and Satyarth Vaidya
- A Deep Reinforced Model for Zero-Shot Cross-Lingual Summarization with Bilingual Semantic Similarity Rewards
  Zi-Yi Dou, Sachin Kumar, and Yulia Tsvetkov
- A Question Type Driven and Copy Loss Enhanced Frameworkfor Answer-Agnostic Neural Question Generation
  Xiuyu Wu, Nan Jiang, and Yunfang Wu
- When and Why is Unsupervised Neural Machine Translation Useless?
  Yunsu Kim, Miguel Graça, and Hermann Ney
- A Generative Approach to Titling and Clustering Wikipedia Sections
  Anjalie Field, Sascha Rothe, Simon Baumgartner, Cong Yu, and Abe Ittycheriah
- The Unreasonable Volatility of Neural Machine Translation Models
  Marzieh Fadaee and Christof Monz
- Leveraging Sentence Similarity in Natural Language Generation: Improving Beam Search using Range Voting
  Sebastian Borgeaud and Guy Emerson
- Transformers without Tears: Improving the Normalization of Self-Attention
  Toan Q. Nguyen and Julian Salazar
- Masked Language Model Scoring
  Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff
- Distill, Adapt, Distill: Training Small, In-Domain Models for Neural Machine Translation
  Mitchell Gordon and Kevin Duh
- Improving Neural Machine Translation Using Energy-Based Models
  Subhajit Naskar, Amirmohammad Rooshenas, and Andrew McCallum
- Training and Inference Methods for High-Coverage Neural Machine Translation
  Michael Yang, Yixin Liu, and Rahul Mayuranath
- Meeting the 2020 Duolingo Challenge on a Shoestring
  Tadashi Nomoto
- English-to-Japanese Diverse Translation by Combining Forward and Backward Outputs
  Masahiro Kaneko, Aizhan Imankulova, Tosho Hirasawa, and Mamoru Komachi
- POSTECH Submission on Duolingo Shared Task
  Junsu Park, Hongseok Kwon, and Jong-Hyeok Lee
- The ADAPT System Description for the STAPLE 2020 English-to-Portuguese Translation Task
  Rejwanul Haque, Yasmin Moslem, and Andy Way
- Expand and Filter: CUNI and LMU Systems for the WNGT 2020 Duolingo Shared Task
  Jindřich Libovický, Zdeněk Kasner, Jindřich Helcl, and Ondřej Dušek
- Exploring Model Consensus to Generate Translation Paraphrases
  Zhenhao Li, Marina Fomicheva, and Lucia Specia
- Growing Together: Modeling Human Language Learning With n-Best Multi-Checkpoint Machine Translation
  El Moatez Billah Nagoudi, Muhammad Abdul-Mageed, and Hasan Cavusoglu

- Generating Diverse Translations via Weighted Fine-tuning and Hypotheses Filtering for the Duolingo STAPLE Task
  Sweta Agrawal and Marine Carpuat
- The JHU Submission to the 2020 Duolingo Shared Task on Simultaneous Translation and Paraphrase for Language Education
  Huda Khayrallah, Jacob Bremerman, Arya D. McCarthy, Kenton Murray, Winston Wu, and Matt Post
- Simultaneous paraphrasing and translation by fine-tuning Transformer models
  Rakesh Chada
- The NiuTrans System for WNGT 2020 Efficiency Task
  Chi Hu, Bei Li, Yinqiao Li, Ye Lin, Yanyang Li, Chenglong Wang, Tong Xiao, and Jingbo Zhu
- Efficient and High-Quality Neural Machine Translation with OpenNMT
  Guillaume Klein, Dakun Zhang, Clément Chouteau, Josep Crego, and Jean Senellart
- Edinburgh's Submissions to the 2020 Machine Translation Efficiency Task
  Nikolay Bogoychev, Roman Grundkiewicz, Alham Fikri Aji, Maximiliana Behnke, Kenneth Heafield, Sidharth Kashyap, Emmanouil-Ioannis Farsarakis, and Mateusz Chudyk
- Improving Document-Level Neural Machine Translation with Domain Adaptation
  Sami Ul Haq, Sadaf Abdul Rauf, Arslan Shoukat, and Noor-e- Hira
- Simultaneous Translation and Paraphrase for Language Education
  Stephen Mayhew, Klinton Bicknell, Chris Brust, Bill McDowell, Will Monroe, and Burr Settles

BEA: The 15th Workshop on Innovative Use of NLP for Building Educational Applications

Organizers: *Ekaterina Kochmar, Jill Burstein, Claudia Leacock, Nitin Madnani, Ildiko Pilan, Helen Yannakoudakis, and Torsten Zesch*

[Virtual Portal Website] [Workshop Website]

The schedule here may not be up-to-date, please check the workshop website for the latest one.

- Linguistic Features for Readability Assessment
  Tovly Deutsch, Masoud Jasbi, and Stuart Shieber
- Using PRMSE to evaluate automated scoring systems in the presence of label noise
  Anastassia Loukina, Nitin Madnani, Aoife Cahill, Lili Yao, Matthew S. Johnson, Brian Riordan, and Daniel F. McCaffrey
- Multiple Instance Learning for Content Feedback Localization without Annotation
  Scott Hellman, William Murray, Adam Wiemerslage, Mark Rosenstein, Peter Foltz, Lee Becker, and Marcia Derr
- Complementary Systems for Off-Topic Spoken Response Detection
  Vatsal Raina, Mark Gales, and Kate Knill
- CIMA: A Large Open Access Dialogue Dataset for Tutoring
  Katherine Stasaski, Kimberly Kao, and Marti A. Hearst
- Becoming Linguistically Mature: Modeling English and German Children's Writing Development Across School Grades
  Elma Kerz, Yu Qiao, Daniel Wiechmann, and Marcus Ströbel
- Annotation and Classification of Evidence and Reasoning Revisions in Argumentative Writing
  Tazin Afrin, Elaine Lin Wang, Diane Litman, Lindsay Clare Matsumura, and Richard Correnti
- Can Neural Networks Automatically Score Essay Traits?
  Sandeep Mathias and Pushpak Bhattacharyya
- Tracking the Evolution of Written Language Competence in L2 Spanish Learners
  Alessio Miaschi, Sam Davidson, Dominique Brunato, Felice Dell'Orletta, Kenji Sagae, Claudia Helena Sanchez-Gutierrez, and Giulia Venturi
- Distractor Analysis and Selection for Multiple-Choice Cloze Questions for Second-Language Learners
  Lingyu Gao, Kevin Gimpel, and Arnar Jensson
- Assisting Undergraduate Students in Writing Spanish Methodology Sections
  Samuel González-López, Steven Bethard, and Aurelio Lopez-Lopez
- Applications of Natural Language Processing in Bilingual Language Teaching: An Indonesian-English Case Study
  Zara Maxwelll-Smith, Simón González Ochoa, Ben Foley, and Hanna Suominen
- An empirical investigation of neural methods for content scoring of science explanations
  Brian Riordan, Sarah Bichler, Allison Bradford, Jennifer King Chen, Korah Wiley, Libby Gerard, and Marcia C. Linn
- An Exploratory Study of Argumentative Writing by Young Students: A transformer-based Approach
  Debanjan Ghosh, Beata Beigman Klebanov, and Yi Song
- Should You Fine-Tune BERT for Automated Essay Scoring?
  Elijah Mayfield and Alan W Black
- GECToR — Grammatical Error Correction: Tag, Not Rewrite
  Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi
- Interpreting Neural CWI Classifiers' Weights as Vocabulary Size
  Yo Ehara
- Automated Scoring of Clinical Expressive Language Evaluation Tasks
  Yiyi Wang, Emily Prud'hommeaux, Meysam Asgari, and Jill Dolata
- Context-based Automated Scoring of Complex Mathematical Responses
  Aoife Cahill, James H Fife, Brian Riordan, Avijit Vajpayee, and Dmytro Galochkin
- Predicting the Difficulty and Response Time of Multiple Choice Questions Using Transfer Learning
  Kang Xue, Victoria Yaneva, Christopher Runyon, and Peter Baldwin
- A Comparative Study of Synthetic Data Generation Methods for Grammatical Error Correction
  Max White and Alla Rozovskaya

SIGMORPHON: 17th Workshop on Computational Research in Phonetics, Phonology, and Morphology

Organizers: *Garrett Nicolai and Kyle Gorman*

[Virtual Portal Website] [Workshop Website]

The schedule here may not be up-to-date, please check the workshop website for the latest one.

Friday, July 10

15:30–17:30 **Morning Session**

15:30–16:30 Invited Talk: On Understanding Character-level Models for Representing Morphology (Clara Vania, NYU)

16:30–17:30 Invited Talk: Recursive Schemes for Phonological Analysis (Jane Chandlee, Haverford College)

17:30–17:45 **Break #1**

17:45–19:30 **Shared Task**

17:45–18:00 SIGMORPHON 2020 Shared Task 0: Typologically Diverse Morphological Inflection
Ekaterina Vylomova, Jennifer White, Elizabeth Salesky, et al.

18:00–18:15 The SIGMORPHON 2020 Shared Task on Multilingual Grapheme-to-Phoneme Conversion
Kyle Gorman, Lucas F.E. Ashby, Aaron Goyzueta, Arya McCarthy, Shijie Wu, and Daniel You

18:15–18:30 The SIGMORPHON 2020 Shared Task on Unsupervised Morphological Paradigm Completion
Katharina Kann, Arya D. McCarthy, Garrett Nicolai, and Mans Hulden

18:30–19:30 Shared Task Poster Session (Multiple)

- One-Size-Fits-All Multilingual Models
  Ben Peters and André F. T. Martins
- Ensemble Self-Training for Low-Resource Languages: Grapheme-to-Phoneme Conversion and Morphological Inflection
  Xiang Yu, Ngoc Thang Vu, and Jonas Kuhn
- The CMU-LTI submission to the SIGMORPHON 2020 Shared Task 0: Language-Specific Cross-Lingual Transfer
  Nikitha Murikinati and Antonios Anastasopoulos
- Grapheme-to-Phoneme Conversion with a Multilingual Transformer Model
  Omnia ElSaadany and Benjamin Suter
- The NYU-CUBoulder Systems for SIGMORPHON 2020 Task 0 and Task 2
  Assaf Singer and Katharina Kann
- The IMS—CUBoulder System for the SIGMORPHON 2020 Shared Task on Unsupervised Morphological Paradigm Completion
  Manuel Mager and Katharina Kann
- SIGMORPHON 2020 Task 0 System Description: ETH Zürich Team
  Martina Forster and Clara Meister
- KU-CST at the SIGMORPHON 2020 Task 2 on Unsupervised Morphological Paradigm Completion
  Manex Agirrezabal and Jürgen Wedekind
- Low-Resource G2P and P2G Conversion with Synthetic Training Data
  Bradley Hauer, Amir Ahmad Habibi, Yixing Luan, Arnob Mallik, and Grzegorz Kondrak
- Frustratingly Easy Multilingual Grapheme-to-Phoneme Conversion
  Nikhil Prabhu and Katharina Kann
- Exploring Neural Architectures And Techniques For Typologically Diverse Morphological Inflection
  Pratik Jayarao, Siddhanth Pillay, Pranav Thombre, and Aditi Chaudhary
- University of Illinois Submission to the SIGMORPHON 2020 Shared Task 0: Typologically Diverse Morphological Inflection
  Marc Canby, Aidana Karipbayeva, Bryan Lunt, Sahand Mozaffari, Charlotte Yoder, and Julia Hockenmaier
- One Model to Pronounce Them All: Multilingual Grapheme-to-Phoneme Conversion With a Transformer Ensemble
  Kaili Vesik, Muhammad Abdul-Mageed, and Miikka Silfverberg
- Leveraging Principal Parts for Morphological Inflection
  Ling Liu and Mans Hulden

- Linguist vs. Machine: Rapid Development of Finite-State Morphological Grammars
  Sarah Beemer, Zak Boston, April Bukoski, Daniel Chen, Princess Dickens, Andrew Gerlach, Torin Hopkins, Parth Anand Jawale, Chris Koski, Akanksha Malhotra, Piyush Mishra, Saliha Muradoglu, Lan Sang, Tyler Short, Sagarika Shreevastava, Elizabeth Spaulding, Testumichi Umada, Beilei Xiang, Changbing Yang, and Mans Hulden
- CLUZH at SIGMORPHON 2020 Shared Task on Multilingual Grapheme-to-Phoneme Conversion
  Peter Makarov and Simon Clematide
- The UniMelb Submission to the SIGMORPHON 2020 Shared Task 0: Typologically Diverse Morphological Inflection
  Andreas Scherbakov
- Data Augmentation for Transformer-based G2P
  Zach Ryan and Mans Hulden

19:30–21:00 **Lunch**

21:00–22:06 **Paper Session**

21:00–21:10 Transliteration for Cross-Lingual Morphological Inflection
Nikitha Murikinati, Antonios Anastasopoulos, and Graham Neubig

21:11–21:21 Evaluating Neural Morphological Taggers for Sanskrit
Ashim Gupta, Amrith Krishna, Pawan Goyal, and Oliver Hellwig

21:22–21:32 Getting the ##life out of living: How Adequate Are Word-Pieces for Modelling Complex Morphology?
Stav Klein and Reut Tsarfaty

21:33–21:43 Induced Inflection-Set Keyword Search in Speech
Oliver Adams, Matthew Wiesner, Jan Trmal, Garrett Nicolai, and David Yarowsky

21:44–21:54 Representation Learning for Discovering Phonemic Tone Contours
Bai Li, Jing Yi Xie, and Frank Rudzicz

21:55–22:05 Joint learning of constraint weights and gradient inputs in Gradient Symbolic Computation with constrained optimization
Max Nelson

22:06–22:30 **Best Paper Session**

22:06–22:18 In search of isoglosses: continuous and discrete language embeddings in Slavic historical phonology
Chundra Cathcart and Florian Wandl

22:18–22:30 Multi-Tiered Strictly Local Functions
Phillip Burness and Kevin McMullin

22:30–23:00 **Break #2**

23:00–1:00 (+1) **Afternoon Session**

23:00–0:00 (+1) Invited Talk: Inflectional data science and human/computer-aided linguistic analysis (Robert Malouf, San Diego State University)

Invited Talk: Modeling failure in morphophonological learning (Bruce Hayes, UCLA)

NLPMC: NLP for Medical Conversations

Organizers: *Parminder Bhatia, Chaitanya Shivade, Mona Diab, Byron Wallace, Rashmi Gangadharaiah, Nan Du, Izhak Shafran, and Steven Lin*

[Virtual Portal Website] [Workshop Website]

The schedule here may not be up-to-date, please check the workshop website for the latest one.

Friday, July 10

16:00–17:30 **Morning Session I**
16:00–16:15 Welcome and Opening Remarks
16:15–17:00 Invited Talk (Tanuj Gupta)
17:00–17:15 Methods for Extracting Information from Messages from Primary Care Providers to Specialists
Xiyu Ding, Michael Barnett, Ateev Mehrotra, and Timothy Miller
17:15–17:30 Towards Understanding ASR Error Correction for Medical Conversations
Anirudh Mani, Shruti Palaskar, and Sandeep Konam

18:00–19:30 **Morning Session II**
18:00–18:45 Invited Talk (Anitha Kannan)
18:45–19:30 Invited Talk (Steven Bendrick)

20:30–22:30 **Afternoon Session I**
20:30–21:15 Invited Talk (Judy Chang)
21:15–21:30 Studying Challenges in Medical Conversation with Structured Annotation
Nan Wang, Yan Song, and Fei Xia
21:30–21:45 Generating Medical Reports from Patient-Doctor Conversations Using Sequence-to-Sequence Models
Seppo Enarvi, Marilisa Amoia, Miguel Del-Agua Teba, Brian Delaney, Frank Diehl, Stefan Hahn, Kristina Harris, Liam McGrath, Yue Pan, Joel Pinto, Luca Rubini, Miguel Ruiz, Gagandeep Singh, Fabian Stemmer, Weiyi Sun, Paul Vozila, Thomas Lin, and Ranjani Ramamurthy
21:45–22:00 Towards an Ontology-based Medication Conversational Agent for PrEP and PEP
Muhammad Amith, Licong Cui, Kirk Roberts, and Cui Tao
22:00–22:15 Heart Failure Education of African American and Hispanic/Latino Patients: Data Collection and Analysis
Itika Gupta, Barbara Di Eugenio, Devika Salunke, Andrew Boyd, Paula Allen-Meares, Carolyn Dickens, and Olga Garcia
22:15–22:30 On the Utility of Audiovisual Dialog Technologies and Signal Analytics for Real-time Remote Monitoring of Depression Biomarkers
Michael Neumann, Oliver Roessler, David Suendermann-Oeft, and Vikram Ramanarayanan

23:00–1:00 (+1) **Afternoon Session II**
23:00–23:45 Invited Talk (Adam Miner)
23:45–0:15 (+1) Lightning Talk

Saturday, July 11

0:15–0:30 Robust Prediction of Punctuation and Truecasing for Medical ASR
Monica Sunkara, Srikanth Ronanki, Kalpit Dixit, Sravan Bodapati, and Katrin Kirchhoff
0:30–0:45 Topic-Based Measures of Conversation for Detecting Mild CognitiveImpairment
Meysam Asgari, Liu Chen, and Hiroko Dodge
0:45–1:00 Closing Remarks

ECNLP: The Third Workshop on e-Commerce and NLP

Organizers: *Shervin Malmasi, Eugene Agichtein, Surya Kallumadi, Oleg Rokhlenko, Nicola Ueffing, and Ido Guy*

[Virtual Portal Website] [Workshop Website]

The schedule here may not be up-to-date, please check the workshop website for the latest one.

Friday, July 10

12:30–12:40 Opening ECNLP - East (Ido Guy)

12:40–13:00 Bootstrapping Named Entity Recognition in E-Commerce with Positive Unlabeled Learning
Hanchu Zhang, Leonhard Hennig, Christoph Alt, Changjian Hu, Yao Meng, and Chao Wang

13:00–13:20 Semi-supervised Category-specific Review Tagging on Indonesian E-Commerce Product Reviews
Meng Sun, Marie Stephen Leo, Eram Munawwar, Paul C. Condylis, Sheng-yi Kong, Seong Per Lee, Albert Hidayat, and Muhamad Danang Kerianto

13:20–13:40 Using Large Pretrained Language Models for Answering User Queries from Product Specifications
Kalyani Roy, Smit Shah, Nithish Pai, Jaidam Ramtej, Prajit Nadkarni, Jyotirmoy Banerjee, Pawan Goyal, and Surender Kumar

13:40–14:00 On Application of Bayesian Parametric and Non-parametric Methods for User Cohorting in Product Search
Shashank Gupta

14:00–14:20 Semi-Supervised Iterative Approach for Domain-Specific Complaint Detection in Social Media
Akash Gautam, Debanjan Mahata, Rakesh Gosangi, and Rajiv Ratn Shah

14:20–14:40 Item-based Collaborative Filtering with BERT
Tian Wang and Yuyangzi Fu

14:40–15:00 Deep Hierarchical Classification for Category Prediction in E-commerce System
Dehong Gao

15:00–15:20 Effective Identification of Distinctive Wordmarks (Sujatha Das Gollapalli)

15:20–16:00 Coffee Break

16:00–16:10 Open ECNLP

16:10–16:30 Deep Learning-based Online Alternative Product Recommendations at Scale
Mingming Guo, Nian Yan, Xiquan Cui, San He Wu, Unaiza Ahsan, Rebecca West, and Khalifeh Al Jadda

16:30–16:50 How to Grow a (Product) Tree: Personalized Category Suggestions for eCommerce Type-Ahead
Jacopo Tagliabue, Bingqing Yu, and Marie Beaulieu

16:50–17:40 Keynote 1 (Julian McAuley)

17:40–18:00 Coffee Break

18:00–18:30 Keynote 1 (Heng Ji)

18:30–18:50 Study on Price Consistency regarding Pack Size via Product Variant Retrieval and Pack Size Extraction (Yang Liu)

18:50–19:10 A Deep Learning System for Sentiment Analysis of Service Calls
Yanan Jia

19:10–19:30 Improving Intent Classification in an E-commerce Voice Assistant by Using Inter-Utterance Context
Arpit Sharma

20:30–20:50 SimsterQ: A Similarity based Clustering Approach to Opinion Question Answering
Aishwarya Ashok, Ganapathy Natarajan, Ramez Elmasri, and Laurel Smith-Stvan

20:50–21:10 e-Commerce and Sentiment Analysis: Predicting Outcomes of Class Action Lawsuits
Stacey Taylor and Vlado Keselj

21:10–21:15 Closing ECNLP

SocialNLP: The Eighth International Workshop on Natural Language Processing for Social Media

Organizers: *Lun-Wei Ku and Cheng-Te Li*

[Virtual Portal Website] [Workshop Website]

The schedule here may not be up-to-date, please check the workshop website for the latest one.

Friday, July 10

16:05–16:10 Opening

16:10–17:10 Keynote: Managing Information and Debunking Misinformation in the Time of Covid-19 (Pascale Fung, Hong Kong University of Science and Technology)

17:30–17:50 Coffee Break

17:50–18:30 **Technical Session 1**

17:50–18:10 Enhancing Bias Detection in Political News Using Pragmatic Presupposition
Lalitha Kameswari, Dama Sravani, and Radhika Mamidi

18:10–18:30 Demoting Racial Bias in Hate Speech Detection
Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov

18:30–19:30 **Data Session**

18:30–18:50 NARMADA: Need and Available Resource Managing Assistant for Disasters and Adversities
Kaustubh Hiware, Ritam Dutt, Sayan Sinha, Sohan Patro, Kripa Ghosh, and Saptarshi Ghosh

18:50–19:10 BEEP! Korean Corpus of Online News Comments for Toxic Speech Detection
Jihyung Moon, Won Ik Cho, and Junbum Lee

19:10–19:30 Stance Prediction for Contemporary Issues: Data and Experiments
Marjan Hosseinia, Eduard Dragut, and Arjun Mukherjee

19:30–21:00 Lunch

21:00–22:00 EmotionGIF Challenge

22:00–22:30 Coffee Break

22:30–22:50 Challenges in Emotion Style Transfer: An Exploration with a Lexical Substitution Pipeline
David Helbig, Enrica Troiano, and Roman Klinger

22:50–23:10 Incorporating Uncertain Segmentation Information into Chinese NER for Social Media Text
Shengbin Jia, Ling Ding, Xiaojun Chen, Shijia E, and Yang Xiang

23:10–23:30 Multi-Task Supervised Pretraining for Neural Domain Adaptation
Sara Meftah, Nasredine Semmar, Mohamed-Ayoub Tahiri, Youssef Tamaazousti, Hassane Essafi, and Fatiha Sadat

23:30–23:35 **Closing Remark**

AutoSimTrans: The 1st Workshop on Automatic Simultaneous Translation: challenges, recent advances,and future directions

Organizers: *Hua Wu, Colin Cherry, James Cross, Liang Huang, Zhongjun He, Mark Liberman, and Yang Liu*

[Virtual Portal Website] [Workshop Website]

The schedule here may not be up-to-date, please check the workshop website for the latest one.

Friday, July 10

14:20–14:30 Opening Remarks

14:30–16:30 **Session 1**
14:30–15:00 Invited Talk 1: Colin Cherry
15:00–15:30 Invited Talk 2: Barry Slaughter Olsen
15:30–16:00 Invited Talk 3: Jordan Boyd-Graber
16:00–16:30 Q&A

16:30–22:00 **Break**

22:00–23:10 **Session 2: Research Paper and System Description**
22:00–22:10 Dynamic Sentence Boundary Detection for Simultaneous Translation
Ruiqing Zhang and Chuanqiang Zhang
22:10–22:20 End-to-End Speech Translation with Adversarial Training
Xuancai Li, Chen Kehai, Tiejun Zhao, and Muyun Yang
22:20–22:30 Robust Neural Machine Translation with ASR Errors
Haiyang Xue, Yang Feng, Shuhao Gu, and Wei Chen
22:30–22:40 Improving Autoregressive NMT with Non-Autoregressive Model
Long Zhou, Jiajun Zhang, and Chengqing Zong
22:40–22:50 Modeling Discourse Structure for Document-level Neural Machine Translation
Junxuan Chen, Xiang Li, Jiarui Zhang, Chulun Zhou, Jianwei Cui, Bin Wang, and Jinsong Su
22:50–23:00 BIT's system for the AutoSimTrans 2020
Minqin Li, Haodong Cheng, Yuanjie Wang, Sijia Zhang, Liting Wu, and Yuhang Guo
23:00–23:10 Q&A

23:10–23:20 **Break**

23:20–1:20 (+1) **Session 3**
23:20–23:50 Invited Talk 4: Hua Wu
23:50–0:20 (+1) Invited Talk 5: Kay-Fan Cheung

Saturday, July 11

0:20–0:50 Invited Talk 6: Qun Liu
0:50–1:20 Q&A

1:20–1:30 **Closing Remarks**

Challenge-HML: The Second Grand-Challenge and Workshop on Human Multimodal Language

Organizers: *AmirAli Bagher Zadeh, Louis-Philippe Morency, Paul Pu Liang, Soujanya Poria, and Ying Shen*

[Virtual Portal Website] [Workshop Website]

The schedule here may not be up-to-date, please check the workshop website for the latest one.

- A Transformer-based joint-encoding for Emotion Recognition and Sentiment Analysis
  Jean-Benoit Delbrouck, Noé Tits, Mathilde Brousmiche, and Stéphane Dupont
- A Multi-modal Approach to Fine-grained Opinion Mining on Video Reviews
  Edison Marrese-Taylor, Cristian Rodriguez, Jorge Balazs, Stephen Gould, and Yutaka Matsuo
- Multilogue-Net: A Context-Aware RNN for Multi-modal Emotion Detection and Sentiment Analysis in Conversation
  Aman Shenoy and Ashish Sardana
- Low Rank Fusion based Transformers for Multimodal Sequences
  Saurav Sahay, Eda Okur, shachi H Kumar, and Lama Nachman
- Unsupervised Online Grounding of Natural Language during Human-Robot Interactions
  Oliver Roesler
- Leveraging Multimodal Behavioral Analytics for Automated Job Interview Performance Assessment and Feedback
  Anumeha Agrawal, Rosa Anil George, Selvan Sunitha Ravi, Sowmya Kamath S, and Anand Kumar
- Audio-Visual Understanding of Passenger Intents for In-Cabin Conversational Agents
  Eda Okur, shachi H Kumar, Saurav Sahay, and Lama Nachman
- AI Sensing for Robotics using Deep Learning based Visual and Language Modeling
  yuvaram singh yuvaram and Kameshwar Rao JV
- Exploring Weaknesses of VQA Models through Attribution Driven Insights
  Shaunak Halbe

7

Anti-harassment Policy

The open exchange of ideas, the freedom of thought and expression, and respectful scientific debate are central to the aims and goals of the ACL. These require a community and an environment that recognizes the inherent worth of every person and group, that fosters dignity, understanding, and mutual respect, and that embraces diversity. For these reasons, ACL is dedicated to providing a harassment-free experience for all the members, as well as participants at our events and in our programs.

Harassment and hostile behavior are unwelcome at any ACL conference, associated event, or in ACL-affiliated online discussions. This includes: speech or behavior that intimidates, creates discomfort, or interferes with a person's participation or opportunity for participation in a conference or an event. We aim for ACL-related activities to be an environment where harassment in any form does not happen, including but not limited to: harassment based on race, gender, religion, age, color, appearance, national origin, ancestry, disability, sexual orientation, or gender identity. Harassment includes degrading verbal comments, deliberate intimidation, stalking, harassing photography or recording, inappropriate physical contact, and unwelcome sexual attention. The policy is not intended to inhibit challenging scientific debate, but rather to promote it through ensuring that all are welcome to participate in shared spirit of scientific inquiry. Vexatious complaints and willful misuse of this procedure will render the complainant subject to the same sanctions as a violation of the anti-harassment policy.

It is the responsibility of the community as a whole to promote an inclusive and positive environment for our scholarly activities. In addition, anyone who experiences harassment or hostile behavior may contact any current member of the ACL Executive Committee or contact Priscilla Rasmussen (acl@aclweb.org), who is usually available at the registration desk during ACL conferences. Members of the executive committee will be instructed to keep any such contact in strict confidence, and those who approach the committee will be consulted before any actions are taken.

The ACL Executive Committee members are listed at:

`https://www.aclweb.org/portal/about`

The full policy and its implementation is defined at:

`https://www.aclweb.org/adminwiki/index.php?title=Anti-Harassment_Policy`

Approved by ACL Executive Committee, 2016

Revised by ACL Executive Committee, July 15, 2018

8 ACL Author Guidelines

Preserving Double Blind Review

The following rules and guidelines are meant to protect the integrity of double-blind review and ensure that submissions are reviewed fairly. The rules make reference to the anonymity period, which runs from 1 month before the submission deadline up to the date when your paper is either accepted, rejected, or withdrawn.

- You may not make a non-anonymized version of your paper available online to the general community (for example, via a preprint server) during the anonymity period. By a version of a paper we understand another paper having essentially the same scientific content but possibly differing in minor details (including title and structure) and/or in length (e.g., an abstract is a version of the paper that it summarizes).

- If you have posted a non-anonymized version of your paper online before the start of the anonymity period, you may submit an anonymized version to the conference. The submitted version must not refer to the non-anonymized version, and you must inform the program chair(s) that a non-anonymized version exists. You may not update the non-anonymized version during the anonymity period, and we ask you not to advertise it on social media or take other actions that would further compromise double-blind reviewing during the anonymity period.

- Note that, while you are not prohibited from making a non-anonymous version available online before the start of the anonymity period, this does make double-blind reviewing more difficult to maintain, and we therefore encourage you to wait until the end of the anonymity period if possible. Alternatively, you may consider submitting your work to the Computational Linguistics journal, which does not require anonymization and has a track for "short" (i.e., conference-length) papers.

Citation and Comparison

If you are aware of previous research that appears sound and is relevant to your work, you should cite it even if it has not been peer-reviewed, and certainly if it influenced your own work. However, refereed publications take priority over unpublished work reported in preprints. Specifically:

- You are expected to cite all refereed publications relevant to your submission, but you may be excused for not knowing about all unpublished work (especially work that has been recently posted and/or is not widely cited).
- In cases where a preprint has been superseded by a refereed publication, the refereed publication should be cited in addition to or instead of the preprint version.

Papers (whether refereed or not) appearing less than 3 months before the submission deadline are considered contemporaneous to your submission, and you are therefore not obliged to make detailed comparisons that require additional experimentation and/or in-depth analysis.

Index

, amittai axelrod amittai, 619
, anxiang zhang anxiang, 443, 483
, arthur szlam arthur, 341, 384
, changliang li changliang, 246, 285, 447, 485
, feng wu feng, 271, 332
, hong yu hong, 149, 189
, hui xue hui, 233, 272
, jiaze wei jiaze, 620
, jingfang xu jingfang, 266, 305
, junjie cao junjie, 452, 469
, kobi leins kobi, 224, 281
, liang he liang, 440, 480
, qihang feng qihang, 30, 87
, randy zhong randy, 479, 519
, sebsibe hailemariam sebsibe, 616
, sheng zhao sheng, 33, 128
, wei zou wei, 246, 285
, wenbin liu wenbin, 620
, wenzheng feng wenzheng, 231, 288
, william ferguson william, 621
, yanqing he yanqing, 620
, yongqiang zhao yongqiang, 64, 107
, you pan you, 620
, yuvaram singh yuvaram, 645
, zhenfeng wu zhenfeng, 620
, zijian li zijian, 35, 129
Şahin, Gözde Gül, 75, 116, 349, 390
Žilinec, Matúš, 620

A, kalaivani, 627
Abate, Solomon Teferra, 616
Abdou, Mostafa, 274, 313, 512, 570
Abdul Rauf, Sadaf, 617, 637
Abdul-Mageed, Muhammad, 636, 639
Abe, Kaori, 254, 273
Abend, Omri, 74, 115, 435, 502, 531, 598
Abera, Hafte, 616
Abnar, Samira, 283, 322
Abrego, Gustavo Hernandez, 140, 199
Acharya, Bharvee, 627
Acken, Audrey, 617
Adams, Douglas, 373, 392
Adams, Oliver, 640
Adel, Heike, 75, 116, 467, 506, 632, 633
Adhikari, Ashutosh, 633, 634
Afrin, Tazin, 638
Agarwal, Rohit, 627
Agarwal, Sanchit, 621
Agarwal, Shubham, 547, 585
Agarwal, Sumeet, 294, 312
Aggarwal, Samarth, 423, 460
Agić, Željko, 157, 196
Agichtein, Eugene, 642
Agirre, Eneko, 54, 72, 294, 312, 488, 492, 528, 531
Agirrezabal, Manex, 639
Agrawal, Ameeta, 374, 393
Agrawal, Anumeha, 645
Agrawal, Sweta, 617, 637
Aguilar, Gustavo, 543, 598
Aharoni, Roee, 526, 563
Aharonov, Ranit, 471, 513
Ahmad, Amin, 140, 199
Ahmad, Wasi, 352, 411
Ahrens, Kathleen, 627
Ahsan, Unaiza, 642
Ahuja, Kabir, 500, 542
Ahuja, Ojas, 621
Ai, Di, 619
Aizawa, Akiko, 69, 112
Aji, Alham Fikri, 525, 562, 636, 637
Akama, Reina, 47, 86
Akdemir, Arda, 58, 472
Akinfaderin, Adewale, 617
Akula, Arjun, 446, 586
Akyürek, Afra Feyza, 566, 603
Al Jadda, Khalifeh, 642
Al Khatib, Khalid, 74, 115, 232, 270, 471, 513
Al-Onaizan, Yaser, 352, 411, 446, 562, 586, 587, 620
Alayrac, Jean-Baptiste, 167, 207
Aldawsari, Mohammed, 629
Alemahu, Ribka, 616
Alemayehu, Ribka, 616
Alemneh, Girma Neshir, 616
Aletras, Nikolaos, 299, 361

Alhindi, Tariq, 565, 602
Ali, Ahmed, 242, 260
Ali, Wazir, 51, 111
Alikhani, Malihe, 11, 19, 446, 484
Alipoormolabashi, Pegah, 616
Alishahi, Afra, 29, 203, 282, 301
Aljunied, Sharifah Mahani, 423, 461
Alkhouli, Tamer, 619, 620
Allen, Carl, 623
Allen-Meares, Paula, 641
Almonte, Andy, 515, 593
Alnafesah, Ghadi, 628
Alqahtani, Sawsan, 549, 604
Alshomary, Milad, 291, 311
Alsudais, Abdulkareem, 631
Alt, Christoph, 90, 91, 98, 130, 131, 431, 642
Aluísio, Sandra, 616
Alva-Manchego, Fernando, 330, 348
Amigo, Enrique, 267, 371
Amin, Saadullah, 621, 623
Amith, Muhammad, 641
Ammar, Waleed, 535, 576
Amoia, Marilisa, 641
Amplayo, Reinald Kim, 138, 178
An, Aijun, 374, 393
An, Bang, 68, 129
An, Jisun, 242, 260
Anand Jawale, Parth, 640
Anand, Tanvi, 617
Anand, Vivek, 627
Ananiadou, Sophia, 502, 583, 622
Anastasopoulos, Antonios, 93, 133, 566, 603, 639, 640
Anderson, Mark, 625, 626
Anderson, Peter, 631
Anderson, Tim, 619
Andreas, Jacob, 509, 588
Andrews, Nicholas, 544, 599, 633, 634
Androutsopoulos, Ion, 289, 329, 623
Anees, Yusra, 617
Angelidis, Stefanos, 374, 393
Angelov, Krasimir, 274, 313
Anil George, Rosa, 645
Ansari, Ebrahim, 619
Ao, Xiang, 549, 604
Araki, Jun, 149, 189
Arase, Yuki, 36, 70
Araujo, Vladimir, 616, 617
Arivazhagan, Naveen, 173, 192, 620
Arkhangorodsky, Arkady, 619
Armendariz, Carlos Santos, 29, 85
Arnold, Matt, 315, 377
Arora, Aryaman, 527, 550
Arora, Kushal, 175, 215
Arora, Simran, 169, 209
Artetxe, Mikel, 294, 312, 321, 367, 492, 531
Artzi, Yoav, 167, 169, 207, 209
Arviv, Ofir, 74, 115
Asahara, Masayuki, 632, 634
Asai, Akari, 368, 388
Asente, Paul, 565, 602
Asgari, Meysam, 638, 641
Ashby, Lucas F.E., 639
Ashok, Aishwarya, 642
Assabie, Yaregal, 616
Atanasov, Atanas, 46, 64
Atanasova, Pepa, 490, 569
Atinafu, Solomon, 616
Atkins, David, 253, 410
Atnafu, Solomon, 616
Atrasevych, Vitaliy, 638
Attardi, Giuseppe, 626
Augenstein, Isabelle, 486, 490, 527, 569, 632
August, Tal, 629
Augustyniak, Lukasz, 617
Aulamo, Mikko, 256, 337, 619
Auli, Michael, 173, 192
Averbuch-Elor, Hadar, 167, 207
Avvaru, Adithya, 627
Awa, Emmanuel, 179, 216
Awasthi, Abhijeet, 632, 633
Awiszus, Maximilian, 619
Aziz, Wilker, 485, 504
Azzam, Muhammad, 617

Börschinger, Benjamin, 492, 532
Bücker, Sebastian, 88, 127
Bañón, Marta, 307, 350
Babafemi, Oyinlola, 617
Babanejad, Nastaran, 374, 393
Babulkov, Nikolay, 248, 330
Bach, Nguyen, 237, 276, 619
Bagher Zadeh, AmirAli, 157, 196
Bahar, Parnia, 265, 304, 619
Baheti, Ashutosh, 33, 67
Bahri, Dara, 35, 186
Bai, Bing, 281, 319
Bai, Ke, 262, 320
Bai, Kun, 281, 319
Bai, Xuanyu, 54, 72
Bailey, Peter, 353, 412
Bailly, Raphaël, 40, 198
Bak, JinYeong, 441, 520
Bakovic, Eric, 146, 204
Balažević, Ivana, 623
Balasubramanian, Aruna, 305, 388
Balasubramanian, Niranjan, 305, 331, 361, 381, 388, 392
Balasubramanian, Sriram, 632, 633
Balazs, Jorge, 645
Baldridge, Jason, 547, 585, 631
Baldwin, Peter, 638
Baldwin, Timothy, 224, 281, 350, 390, 622, 623
Bali, Kalika, 435, 492
Balkir, Esma, 625
Ballah, Deddeh, 622
Baly, Ramy, 242, 260
Bambrick, Joshua, 515, 593
Banerjee, Jyotirmoy, 642
Banerjee, Souvik, 632, 633
Banisakher, Deya, 629
Bansal, Hritik, 294, 312
Bansal, Mohit, 167, 207, 345, 348, 366, 385, 404, 408, 547, 570, 586, 591, 624
Bansal, Srijan, 64, 107
Bao, Siqi, 31, 109
Bao, Yu, 50, 89
Bapna, Ankur, 173, 192
Bar-Haim, Roy, 270, 331
Barbhuiya, Ferdous, 627
Bareket, Dan, 492, 531
Barkan, Oren, 263, 324
Barnes, Jeremy, 90, 130
Barnett, Michael, 641
Barocas, Solon, 364, 401
Baroni, Marco, 243, 301, 322, 342
Barrón-Cedeño, Alberto, 418, 515
Barra-Chicote, Roberto, 620
Barrett, Maria, 512, 570
Barrow, Joe, 36, 111, 260, 300
Barry, James, 573, 608, 626
Baruah, Arup, 627
Basili, Roberto, 149, 189
Basta, Christine, 617

Bastan, Mohaddeseh, 331, 392
Bastianelli, Emanuele, 523, 547
Bauer, Lisa, 624
Baumgartner, Simon, 636
Bean, Daniel, 622
Beaulieu, Marie, 642
Becker, Lee, 638
Becker, Markus, 549, 604
Bedrick, Steven, 617
Beemer, Sarah, 640
Behnke, Maximiliana, 637
Behnke, Sven, 90, 130
Beigman Klebanov, Beata, 531, 609, 627, 628, 638
Bekki, Daisuke, 312, 429, 490, 530
Belding, Elizabeth, 225, 281
Belgrano, Lorenzo, 157, 196
Belinkov, Yonatan, 11, 14, 133, 174, 322, 343, 512, 569, 570, 590, 621
Beltagy, Iz, 152, 213, 502, 551, 599, 606
Belyy, Anton, 629
Ben Abacha, Asma, 631
Benamara, Farah, 271, 331
Bendayan, Rebecca, 622
Bender, Emily M., 11, 21, 355, 394
Bengio, Yoshua, 375, 447, 504, 592
Benotti, Luciana, 616, 631
Benteau, Renou, 75, 116
Bentivogli, Luisa, 465, 562
Berant, Jonathan, 55, 72, 305, 367, 368, 386
Berg, Tamara, 167, 207, 547, 586
Berg-Kirkpatrick, Taylor, 226, 346, 484, 549, 586, 604
Bergen, Leon, 146, 204
Bernaczyk, Michał, 617
Bernard, Timothée, 60, 139
Bernardi, Raffaella, 616, 631
Bernardy, Jean-Philippe, 626
Berndt, Jakob, 97, 117
Berzak, Yevgeni, 371, 409
Besacier, Laurent, 619
Bethard, Steven, 306, 342, 386, 389, 559, 582, 622, 638
Betke, Margrit, 566, 603
Bevendorff, Janek, 74, 115
Bevilacqua, Michele, 175, 215, 330, 349
Beyene, Million Meshesha, 616
Beymer, David Beymer, 622
Bhalla, Grusha, 375, 607
Bharadwaj, Akash, 532, 609
Bhargav, G P Shrivatsa, 171, 211
Bhargava, Aditya, 633, 634
Bhargava, Prajjwal, 39, 491
Bhat, Siddharth, 632, 633
Bhat, Suma, 627
Bhathena, Hanoz, 633, 634
Bhatia, Parminder, 641
Bhatt, Gantavya, 294, 312
Bhattacharyya, Pushpak, 292, 293, 333, 334, 620, 629, 638
Bhooshan, Suvrat, 569, 590
Bhutani, Nikita, 635
Bi, Wei, 48, 65, 169, 209, 422, 519
Bianchi, Federico, 94, 114
Bichler, Sarah, 638
Bicknell, Klinton, 637
Biemann, Chris, 226, 284
Biesialska, Magdalena, 312, 572
Bilu, Yonatan, 471, 513
Bin, Yi, 266, 305
Binau, Julie, 624
Bing, Lidong, 35, 68, 423, 461
Birch, Alexandra, 636
Bird, Steven, 449, 486
Bizzoni, Yuri, 620, 627
Björkelund, Anders, 625
Black, Alan W, 128, 158, 171, 186, 197, 211, 307, 371, 488, 528, 638
Blain, Frédéric, 75, 116
Blanco, Eduardo, 551, 552, 589, 606
Blei, David, 362, 382
Bleiweiss, Avi, 623
Blevins, Terra, 57, 175
Blix, Hagen, 449, 486
Blodgett, Su Lin, 364, 401
Blunsom, Phil, 167, 207, 264, 282, 301, 326
Bodapati, Sravan, 641
Bogin, Ben, 367, 386
Bogoychev, Nikolay, 93, 134, 525, 562, 619, 637
Bohnet, Bernd, 137, 178, 444, 483
Bojar, Ondřej, 294, 312, 619, 620
Boleda, Gemma, 282, 302
Bollegala, Danushka, 51, 70
Bollmann, Marcel, 532, 609
Bolton, Jason, 160, 216
Bommasani, Rishi, 343, 386
Bonial, Claire, 629
Bordes, Antoine, 164, 205, 330, 348
Borgeaud, Sebastian, 636
Borgholt, Lasse, 157, 196
Boschee, Elizabeth, 555, 593
Bosselut, Antoine, 11, 20
Bossuyt, Patrick, 622
Boston, Jeff, 315, 377
Boston, Zak, 640
Botner, Nick, 535, 576
Bouchacourt, Diane, 301, 342
Boudin, Florian, 69, 112
Bougares, Fethi, 619
Bouma, Gosse, 625
Boureau, Y-Lan, 147, 164, 205, 206, 341, 384
Bowman, Samuel R., 179, 216, 355, 394, 625
Boyd, Alex, 30, 205
Boyd, Andrew, 641
Boyd-Graber, Jordan, 151, 191, 260, 300, 492, 532
Bražinskas, Arthur, 354, 413
Bradford, Allison, 638
Brandt, Cynthia, 623
Brantley, Kianté, 149, 189
Bredin, Hervé, 632, 633
Bremerman, Jacob, 637
Brignone, Fabrizio, 81, 140
Brix, Christopher, 265, 304
Brockett, Chris, 348, 377, 408, 576
Brooks, Jennifer, 628
Broscheit, Samuel, 155, 194
Brousmiche, Mathilde, 645
Bruijn, Berry de, 623
Brunato, Dominique, 638
Bruni, Elia, 38, 192
Brunk, Clifford, 35, 186
Brusilovsky, Peter, 541, 581
Brust, Chris, 637
Buddemeyer, Amanda, 119, 554
Budhiraja, Amar, 435, 492
Buechel, Sven, 74, 116
Buendia, Alejandro, 635
Bugliarello, Emanuele, 92, 93, 113, 133
Bui, Trung, 547, 585, 635
Buitelaar, Paul, 627
Bukoski, April, 640
Bulat, Luana, 551, 588
Burke, Robert, 499, 541
Burness, Phillip, 640
Burnham, Greg, 532, 609
Burstein, Jill, 638

Burtenshaw, Ben, 628
Bustamante, Gina, 152, 213, 616
Buttery, Paula, 152, 213
Byrne, Bill, 525, 526, 563

C. Linn, Marcia, 638
Côté, Marc-Alexandre, 155, 194
Caciularu, Avi, 263, 324
Cahill, Aoife, 638
Cahyawijaya, Samuel, 252, 333
Cai, Deng, 77, 118
Cai, Han, 525, 601
Cai, Hengyi, 440, 479
Cai, Jiong, 236, 274, 452, 510
Cai, Liwei, 449, 605
Cai, Ruichu, 35, 129
Cai, Yi, 484, 523
Caines, Andrew, 152, 213
Calabrese, Agostina, 330, 349
Calixto, Iacer, 616
Callison-Burch, Chris, 110, 185, 330, 390, 500, 597
Camburu, Oana-Maria, 282, 301
Campagna, Giovanni, 31, 148
Campos, Jon Ander, 488, 528
Canby, Marc, 639
Canny, John, 354, 413, 575, 611
Cao, Guihong, 179, 216
Cao, Jiarun, 444, 461
Cao, Juan, 46, 64
Cao, Jun, 26, 179
Cao, Pengfei, 230, 288, 436, 515
Cao, Qingqing, 305, 388
Cao, Ruisheng, 50, 67, 452, 470
Cao, Yang Trista, 319, 402
Cao, Yifan, 236, 274
Cao, Yixin, 67, 128, 423, 460
Cao, Yuan, 173, 192
Cao, Yue, 432, 473
Caragea, Cornelia, 351, 361, 381, 514
Caragea, Doina, 351, 514
Carbonell, Jaime, 501, 544, 545, 599
Cardie, Claire, 343, 349, 368, 386, 388, 408, 543, 582
Carin, Lawrence, 165, 186, 504, 560
Carmeli, Boaz, 51, 112
Carpuat, Marine, 617, 628, 637
Carrillo-de-Albornoz, Jorge, 267, 371
Carvallo, Andrés, 616
Casanueva, Iñigo, 621
Casas, Noe, 78, 554
Caselli, Tommaso, 629
Casper, Stephen, 621
Castelli, Vittorio, 75, 117, 369, 407
Castellucci, Giuseppe, 149, 189
Caswell, Isaac, 525, 587, 601
Catanzaro, Bryan, 30, 205
Cathcart, Chundra, 640
Cattoni, Roldano, 465, 562, 619
Caubrière, Antoine, 619
Caufield, John, 101, 295
Cavallari, Sandro, 635
Cavusoglu, Hasan, 636
Celikyilmaz, Asli, 348, 408, 621, 631
Cengiz, Cemil, 632, 633
Cer, Daniel, 140, 199
Cerda-Mardini, Patricio, 617
Chaabouni, Rahma, 301, 342
Chada, Rakesh, 637
Chai, Joyce, 5
Chai, Yekun, 464, 504
Chai, Zi, 34, 67
Chakrabarti, Soumen, 423, 460
Chakrabarty, Tuhin, 541, 565, 597, 602
Chakraborty, Aishik, 175, 215
Chakraborty, Saikat, 352, 411
Chakravarti, Rishav, 75, 117, 171, 211
Challis, Christopher, 105, 146
Chambers, Nathanael, 331, 392
Chami, Ines, 464, 505
Chan, Alvin, 362, 382
Chan, Hou Pong, 69, 111
Chan, Iat Chong, 515, 593
Chandel, Shubham, 369, 407
Chandramouli, Rajarathnam, 622
Chang, David, 623
Chang, Ernie, 482, 499
Chang, Jason S., 220, 277
Chang, Kai-Wei, 52, 169, 209, 213, 224, 225, 243, 251, 281, 302, 311, 352, 356, 395, 401, 411
Chang, Ming-Wei, 369, 388, 551, 588
Chang, Nancy, 355, 394
Chang, Shih-Fu, 140, 167, 180, 207
Chang, Tyler A., 633, 634
Chang, Walter, 177, 530
Chang, Xiaojun, 548, 586
Chang, Yi, 90, 130
Chao, Lidia S., 38, 93, 465, 526
Chao, Wenhan, 288, 328
Chaturvedi, Snigdha, 29, 85, 165, 185, 629
Chaudhary, Aditi, 639
Chaudhary, Vishrav, 552, 589
Chaudhury, Sriram, 169, 209
Chauhan, Aabhas, 101, 295
Chauhan, Dushyant Singh, 292, 333
Chauhan, Kushal, 230, 288
Chawla, Daniel, 623
Che, Wanxiang, 31, 42, 65, 86, 108, 120, 124–126, 205, 440, 450, 519, 528
Chemla, Emmanuel, 343, 403
Chen, Alyssa, 617
Chen, Bei, 87, 109
Chen, Boli, 230, 288
Chen, Boxing, 92, 132
Chen, Brian, 140, 180
Chen, Changyou, 51, 68, 111, 129, 165, 186
Chen, Chencai, 440, 480
Chen, Chengbo, 471, 513
Chen, Chengyao, 234, 272
Chen, Daniel, 640
Chen, Danqi, 12, 22, 346, 405
Chen, Daoyuan, 424, 462
Chen, Enhong, 37, 113
Chen, Gang, 424, 461
Chen, Hanjie, 367, 403
Chen, Hannah, 136, 572
Chen, Hong-You, 175, 215
Chen, Hongshen, 440, 479
Chen, Howard, 169, 209
Chen, Hsin-Hsi, 32, 127
Chen, Huajun, 227, 263
Chen, Jhih-Jie, 220, 277
Chen, Jiaao, 150, 190
Chen, Jiacheng, 167, 207
CHEN, Jiajun, 30, 50, 89, 124, 246, 285
Chen, Jianshu, 363, 400, 450, 529, 541, 597
Chen, Jiaoyan, 227, 263
Chen, Jiaze, 26, 179
Chen, Jindong, 621
Chen, Jingmin, 227, 263
Chen, Jingxiang, 621
Chen, Jingya, 2
Chen, Jiun-Hung, 248, 307
Chen, Jiusheng, 451, 488

Chen, Jun, 231, 289
Chen, Junxuan, 644
Chen, Junying, 484, 523
Chen, Ke, 424, 461
Chen, Kehai, 37, 132, 247, 286
Chen, Kunlong, 53, 96
Chen, Li, 26, 179
Chen, Liangyu, 482, 499
Chen, Liu, 641
Chen, Long, 230, 287
Chen, Lu, 50, 67, 430, 452, 470, 511
Chen, Luoxin, 573, 608
Chen, Mia, 173, 192
Chen, Mingda, 633, 634
Chen, Moxin, 30, 124
Chen, Nancy, 362, 382
Chen, Patrick H., 169, 209
Chen, Pinzhen, 93, 134, 307, 350, 619
Chen, Qiaochu, 429, 569
Chen, Qingyu, 623
Chen, Shaowei, 445, 502
Chen, Tengyang, 624
Chen, Tongfei, 559, 571, 591, 598
Chen, Wang, 69, 111
Chen, Wei, 644
Chen, Weizhu, 150, 179, 190, 216
Chen, Wenhu, 33, 185, 541, 597
Chen, Xi (Leslie), 157, 196
Chen, Xianyang, 627–629
Chen, Xiao, 251, 331
Chen, Xiaojun, 643
Chen, Xinlei, 631
Chen, Yao, 35, 129
Chen, Yen-Chun, 377, 540, 576, 580
Chen, Yifu, 30, 87
Chen, Yihong, 87, 109
Chen, Yiran, 432, 473
Chen, Yubo, 230, 288, 436, 515
Chen, Yufei, 452, 509
Chen, Yun, 282, 322
Chen, Yun-Nung, 49, 65, 252, 333
Chen, Yunmo, 559, 598
Chen, Yuxing, 363, 400
Chen, Zhi, 50, 67, 430, 511
Chen, Zhipeng, 42, 120
Chen, Zhiqun, 627
Chen, Zhiyu, 33, 185, 541, 597
Chen, Zhuang, 251, 291
Chen, Zhuohao, 253, 410
Chen, Zixuan, 87, 109
Chen, Ziye, 430, 570
Chen-Burger, Jessica, 90, 130
Cheng, Benny, 260, 300
Cheng, Fei, 335, 514, 632, 634
Cheng, Hao, 179, 216, 369, 388
Cheng, Haodong, 644
Cheng, Hua, 170, 210
Cheng, Jianpeng, 108, 147
Cheng, Meng, 251, 311
Cheng, Minhao, 447, 485
Cheng, Pengyu, 504, 560
Cheng, Shanbo, 93, 133
Cheng, Xingyi, 53, 96
Cheng, Xueqi, 430, 511
Cheng, Yong, 425, 465
Cheng, Yu, 165, 186, 352, 411, 540, 580
Chenthamarakshan, Vijil, 262, 320
Chernodub, Artem, 638
Cherry, Colin, 620, 644
Chersoni, Emmanuele, 627
Cheung, Jackie Chi Kit, 175, 215
Chi, Ethan A., 367, 385
Chi, Ziming, 230, 328, 445, 502
Chinnappa, Dhivya, 551, 606
Chiril, Patricia, 271, 331
Chiruzzo, Luis, 625
Chiu, Billy, 268, 308
Chiu, Ming-Chang, 629
Cho, Hyundong, 164, 205
Cho, Kyunghyun, 341, 352, 384, 411, 632, 633
Cho, Won Ik, 643
Chodroff, Eleanor, 307, 371, 449, 486
Choe, Hyonsu, 625
Choi, Eunsol, 370, 389
Choi, Jinho D., 370, 407, 622, 625, 628
Choi, Jinwook, 98, 453
Choi, Woo Suk, 631
Choi, Yejin, 11, 20, 145, 204, 364, 401
Chollampatt, Shamil, 247, 286
Chong, Weifeng, 230, 288, 436, 515
Choubey, Prafulla Kumar, 363, 400
Choudhury, Monojit, 248, 267, 435, 492
Chouteau, Clément, 637
Chowdhury, Md. Faisal Mahbub, 151, 191
Christodoulopoulos, Christos, 624
Chrupała, Grzegorz, 29, 203, 282, 301
Chu, Chenhui, 36, 70
Chu, Wei, 30, 53, 86, 96
Chu-Carroll, Jennifer, 532, 609
Chua, Tat-Seng, 67, 89, 128, 129
Chuang, Shun-Po, 425, 466
Chudyk, Mateusz, 637
Chung, Tagyoung, 621
Chung, Yi-Ling, 74, 115
Chung, Yu-An, 157, 196
Cieliebak, Mark, 54, 72, 488, 528
Cirik, Volkan, 484, 586
Clark, Elizabeth, 629
Clark, Jonathan H, 370, 389
Clark, Stephen, 167, 207
Clematide, Simon, 487, 527, 640
Clergerie, Éric de la, 485, 560
Clifton, Ann, 616
Climent, Salvador, 628
Cocarascu, Oana, 624
Cocos, Anne, 330, 390
Cohan, Arman, 152, 213, 535, 576
Cohen, Kevin, 11, 17, 622
Cohen, Shay B., 87, 125, 307, 408, 502, 598, 625
Cohen, Trevor, 145, 203, 622
Cohen, Yaara, 622
Cohn, Trevor, 350, 390
Colas, Anthony, 635
Coll-Florit, Marta, 628
Colla, Davide, 508, 568
Collier, Nigel, 97, 117
Collins, Michael, 370, 389
Condylis, Paul C., 642
Conforti, Costanza, 97, 117
Conneau, Alexis, 427, 552, 588, 589
Constant, Noah, 140, 199
Coope, Samuel, 31, 109
Coria, Juan Manuel, 632, 633
Correnti, Richard, 638
Costa, Douglas, 616
Costa-jussà, Marta R., 78, 312, 554, 572, 617
Cotterell, Ryan, 93, 133, 307, 321, 342, 371, 449, 464, 486, 492, 527, 531, 550, 560, 628
Cotterell, Ryan D., 527, 550
Coulomb-Gully, Marlène, 271, 331
Courtland, Maury, 620
Cowen, Alan, 270, 373

Coy, Adam, 622
Craighead, Hannah, 152, 213
Cranenburgh, Andreas van, 232, 291
Crego, Josep, 92, 113, 637
Cremisini, Andres, 629
Creutz, Mathias, 264, 326
Croce, Danilo, 149, 189
Cross, James, 644
Cui, Hongyi, 620
Cui, Jianwei, 620, 644
Cui, Leyang, 87, 108
Cui, Licong, 641
Cui, Xiaohui, 226, 324
Cui, Xiquan, 642
Cui, Yiming, 42, 86, 108, 120
Cui, Zeyu, 36, 69
Culkin, Ryan, 544, 582
Cunha Silva, Rossana da, 616
Cunha, Washington, 546, 584

D, Thenmozhi, 627
D. Havtorn, Jakob, 157, 196
Düwel, Tim, 95, 152, 475, 575
Da San Martino, Giovanni, 248, 330, 418, 515
Dabre, Raj, 335, 514, 636
DADU, TANVI, 627
Daelemans, Walter, 628
Dagan, Gautier, 38, 192
Dagan, Ido, 469, 606
Dagan, Or, 325, 346
Dai, Andrew, 632, 633
Dai, Kuai, 232, 270
Dai, Xiang, 423, 483
Dai, Xiaoya, 231, 289
Dai, Xinyu, 30, 50, 89, 124, 246, 285
Dai, Yi, 443, 501
Dai, Yinpei, 48, 65
Daif, Mahmoud, 235, 530
Dalvi, Fahim, 133, 174, 322, 343, 619
Dana, Saswati, 75, 117
Dandapat, Sandipan, 248, 267
Danescu-Niculescu-Mizil, Cristian, 260, 300, 361, 381
Dang, Dawei, 619
Dankers, Verna, 628
Darwish, Kareem, 46, 64
Das, Dipanjan, 155, 194, 540, 580
Das, Kaushik, 627
Das, Manirupa, 622
Das, Payel, 262, 320
Dash, Sarthak, 151, 191
Dass, Nathan, 633, 634
Datta, Anupam, 342, 386
Daumé III, Hal, 149, 189, 319, 364, 401, 402, 636
Davidson, Sam, 638
Davis, Forrest, 145, 204
Davis, Kelly, 343, 386
Davoodi, Maryam, 362, 382
Dayanik, Erenay, 299, 381
Debnath, Alok, 632, 633
Degen, Judith, 363, 400
Dehghani, Morteza, 364, 401
Dehouck, Mathieu, 626
Del Tredici, Marco, 268, 308
Del-Agua Teba, Miguel, 641
Delaney, Brian, 641
Delbrouck, Jean-Benoit, 645
Dell'Orletta, Felice, 632, 633, 638
Demeter, David, 150, 190
Demner-Fushman, Dina, 622, 631
Demszky, Dorottya, 270, 373, 617
DeNero, John, 92, 113
Deng, Hangyu, 56, 97
Deng, Haotang, 427, 509
Deng, Yao, 620
Deng, Yu, 368, 388
Deng, Zhiwei, 167, 207
Denisov, Pavel, 418, 494
Denton, Emily, 364, 402
Denuyl, Stephen, 364, 402
Derczynski, Leon, 624
Deriu, Jan, 54, 72, 488, 528
Dernoncourt, Franck, 177, 530, 543, 565, 598, 602, 621, 635
Dernoncourt, Frank, 629
Derr, Marcia, 638
Desai, Shrey, 361, 381, 621
Desmond, Michael, 315, 377
Desmulliez, Marc, 90, 130
Deutch, Daniel, 305, 368
Deutsch, Tovly, 621, 638
Dey, Debadeepta, 348, 408
Dey, Kuntal, 627
DeYoung, Jay, 301, 385, 622
Dhingra, Bhuwan, 343, 403
Dhole, Kaustubh, 50, 89
Di Eugenio, Barbara, 630, 641
Di Gangi, Mattia A., 465, 562, 619
Diab, Mona, 352, 411, 549, 565, 602, 604, 621, 641
Dickens, Carolyn, 641
Dickens, Princess, 640
Diehl, Frank, 641
Dillig, Isil, 429, 569
Dinan, Emily, 164, 206, 348, 408
Ding, Chenchen, 40, 99
Ding, Liang, 93, 114
Ding, Ling, 643
Ding, Ning, 69, 111, 449, 486
Ding, Xiyu, 641
Ding, Zixiang, 232, 331
Dinkov, Yoan, 242, 260
Dinu, Georgiana, 562, 587, 620
Dipersio, Christopher, 621
Dixit, Kalpit, 641
Dixon, Lucas, 289, 329
Djokic, Vesna G., 551, 588
Dligach, Dmitriy, 622
Do, Bich-Ngoc, 275, 314
Do, Nam, 627
Dobnik, Simon, 627
Dobson, Richard, 622
Dodge, Hiroko, 641
Dodge, Jesse, 448, 561
Doggett, Erika, 616
Dognin, Pierre, 262, 320
Doitch, Amichay, 170, 210
Dolan, Bill, 348, 357, 377, 408, 576
Dolata, Jill, 638
Donahue, Chris, 165, 185
Dong, Li, 450, 488
Dong, Ning, 562, 587
Dong, Rui, 98, 453
Dong, Xiangjue, 628
Dong, Xin Luna, 11, 15, 544, 559, 583, 599
Dorna, Michael, 175, 215
Dossou, Femi Pancrace Bonaventure, 616
Dou, Dejing, 543, 598
Dou, Zhicheng, 566, 603
Dou, Zi-Yi, 636
Downey, Doug, 150, 152, 190, 213, 551, 606
Draelos, Rachel, 628
Dragut, Eduard, 643
Dredze, Mark, 544, 565, 599, 602, 633, 634

Du, Bo, 284, 325
Du, Chunning, 270, 311
Du, Jiachen, 251, 291
Du, Junping, 432, 473
Du, Nan, 641
Du, Wenyu, 447, 504
Du, Xiaoyu, 480, 520
Du, Xin, 242, 260
Du, Xinya, 543, 582
Dušek, Ondřej, 353, 412, 636
Dua, Dheeru, 54, 154, 368, 388
Duan, Nan, 54, 72, 287, 328, 427, 429, 430, 450, 451, 488, 509, 511, 512, 528
Duan, Xiangyu, 92, 132
Duan, Yu, 34, 110, 230, 287
Dubois, Yann, 38, 192
Duckworth, Daniel, 110, 185
Dudy, Shiran, 617
Duh, Kevin, 238, 243, 315, 321, 436, 495, 633, 634, 636
Dunfield, Katherine Ann, 623
Dunietz, Jesse, 532, 609
Dupont, Stéphane, 645
Dupont, Yoann, 485, 560
Dupoux, Emmanuel, 301, 342
Durme, Benjamin Van, 155, 194
Durmus, Esin, 352, 411
Durrani, Nadir, 133, 174, 322, 343, 619
Durrett, Greg, 34, 166, 429, 569
Dutt, Ritam, 643
Dutta Chowdhury, Koel, 620
Dutta, pratik, 443, 460
Dyer, Chris, 167, 207, 226, 264, 326, 346, 632
Dyer, William, 146, 204
Dziedzic, Adam, 171, 211

E, Shijia, 643
Eavani, Harini, 33, 185
Ebner, Seth, 544, 582
Echevarria, Jose, 565, 602
Eck, Douglas, 110, 185, 500, 597
Eckart de Castilho, Richard, 467, 506
Eden, Lilach, 270, 331
Edin, Joakim, 157, 196
Edunov, Sergey, 173, 192
Eger, Steffen, 79, 93, 133, 138
Ehara, Yo, 638
Ehren, Rafael, 628
Eichenberger, Alexandre, 467, 506
Eisenberg, Joshua, 629
Eisenschlos, Julian, 290, 310
Eisenstein, Jacob, 425, 465
Eisner, Jason, 307, 371
Ek, Adam, 626
Ekbal, Asif, 292, 333
El Baff, Roxanne, 232, 270
Elachqar, Oussama, 165, 186
Elanwar, Randa, 566, 603
Elazar, Yanai, 485, 505
Elbayad, Maha, 619
Elder, Henry, 499, 541
Elfardy, Heba, 621
Elgohary, Ahmed, 148, 184, 260, 300
Elliott, Desmond, 512, 523, 532, 547, 570, 609
Elmasri, Ramez, 642
ElSaadany, Omnia, 639
ElSherief, Mai, 225, 281
Elsner, Micha, 527, 550
Emelianenko, Dmitrii, 132, 173
Emerson, Guy, 268, 308, 493, 532, 636
Emezue, Chris Chinenye, 616
Enarvi, Seppo, 641
Enyedi, Robert, 620
Erdmann, Alexander, 527, 549, 550, 604
Eric, Mihail, 621
Eshetu, Abebawu, 616
Eskander, Ramy, 617
Eskenazi, Maxine, 49, 147, 148, 205
España-Bonet, Cristina, 620
Esplà-Gomis, Miquel, 307, 350
Essafi, Hassane, 643
Essaidi, Farah, 74, 115
Estève, Yannick, 619
Ethayarajh, Kawin, 224, 281
Ettinger, Allyson, 323, 343, 363, 387, 400
Evans, David, 136, 572
Evans, Nicholas, 235, 273
Eyal, Matan, 622

Fabbri, Alexander, 306, 369
Fadaee, Marzieh, 636
Falenska, Agnieszka, 625
Fan, Chuang, 251, 291
Fan, Lu, 66, 126
Fan, Xiaosheng, 432, 473
Fang, Yan, 255, 295
Fang, Yimai, 108, 147
Farghly, Tyler, 31, 109
Farhan, Aamir, 617
Farsarakis, Emmanouil-Ioannis, 637
Faulkner, Adam, 620
Fazel-Zarandi, Maryam, 621
Federico, Marcello, 619, 620
Federmann, Christian, 619
Fei, Hao, 469, 509
Fei, Hongliang, 373, 392
Feldman, Anna, 627
Feldman, Sergey, 152, 213
Felt, Christian, 627
Feng, Jingrong, 423, 501
Feng, Junlan, 480, 520
Feng, Song, 363, 400
Feng, Tiantian, 629
Feng, Yang, 47, 65, 440, 519, 644
Feng, Yansong, 89, 129, 444, 461
Feng, Yue, 544, 582
Feng, Yulan, 147, 205
Feng, Zhangyin, 427, 509
Fern, Xiaoli, 444, 502
Fernández, Raquel, 268, 308
Fernández-González, Daniel, 275, 313, 469, 510
Fernandez Astudillo, Ramón, 128, 165, 369, 407
Ferreira, Deborah, 490, 511
Ferritto, Anthony, 75, 117, 171, 211
Ferrucci, Dave, 532, 609
Fethi, Amal, 74, 115
Field, Anjalie, 636, 643
Fife, James H, 638
Filice, Ross, 137, 178
Filippskikh, Elizaveta, 636
Finch, Andrew, 636
Finegan-Dollak, Catherine, 315, 377
Finlayson, Mark, 629
Firat, Orhan, 173, 192
Flek, Lucie, 532, 609
Flor, Michael, 628, 629
Florian, Radu, 75, 117, 128, 165
Foerster, Jakob, 632, 633
Foley, Ben, 638
Foltz, Peter, 638
Fomicheva, Marina, 75, 116, 636
Fonollosa, José A. R., 78, 554, 617
Fonseca, Erick, 573, 608

Forcada, Mikel L., 307, 350
Fornaciari, Tommaso, 94, 114
Forster, Martina, 639
Fort, Karën, 11, 17
Foryciarz, Agata, 31, 148
Fosler-Lussier, Eric, 622
Foster, George, 620
Foster, Jennifer, 499, 541, 573, 608, 626
Fox Tree, Jean, 29, 85
Frank, Robert, 321, 366
Frank, Stella, 523, 547
Franz, Martin, 75, 117
Fraser, Alexander, 286, 304
Fraser, Kathleen C., 623
Fredrikson, Matt, 342, 386
Freedman, Marjorie, 140, 180
Freitag, Markus, 525, 601
Freitas, André, 490, 511
Freitas, Larissa, 616
Frermann, Lea, 138, 178
Freshia, Sackey, 616
Fried, Daniel, 167, 207
Friedman, Roni, 270, 331
Friedrich, Annemarie, 75, 116, 626
Froes, Jader, 616
Fu, Jie, 155, 194, 362, 382, 451, 488
Fu, Qiankun, 459, 521
Fu, Xiangling, 436, 515
Fu, Yan, 54, 72
Fu, Yuyangzi, 642
Fujinuma, Yoshinari, 151, 191
Fujita, Atsushi, 425, 587, 636
Fukuda, Ryo, 620
Fukumoto, Fumiyo, 627
Fulda, Nancy, 628
Funayama, Hiroaki, 254, 335
Fung, Pascale, 30, 108, 252, 333, 633, 634
Futeral, Matthieu, 74, 115
Futrell, Richard, 146, 204

G'Sell, Maxwell, 226, 346
Gärtner, Markus, 435, 609
Gökçe, Onur, 455, 534
Gábor, Kata, 40, 198
Gómez-Rodríguez, Carlos, 275, 313, 469, 510, 625, 626
Gabriel, Saadia, 364, 401
Gabryszak, Aleksandra, 90, 91, 130, 131
Gaido, Marco, 619
Gales, Mark, 638
Galley, Michel, 357, 377, 576
Gallina, Ygor, 69, 112
Galochkin, Dmytro, 638
Gan, Chuang, 525, 601
Gan, Zhe, 165, 186, 352, 411, 540, 580
Gangadharaiah, Rashmi, 148, 184, 641
Gangal, Varun, 369, 407
Gao, Can, 271, 332
Gao, Dehong, 642
Gao, Jianfeng, 148, 150, 179, 190, 206, 216, 248, 255, 295, 307, 377, 495, 555, 576
Gao, Lingyu, 633, 634, 638
Gao, Shuyang, 621
Gao, Silin, 48, 109
Gao, Tianyu, 443, 501
Gao, Xiang, 357, 377, 576
Gao, Yang, 79, 93, 133, 138
Gao, Yifan, 55, 72
Gao, Yingqiang, 92, 132
Gao, Zhe, 230, 287
Gaonkar, Radhika, 331, 392
Garcia, Olga, 641
Gardner, Matt, 54, 150, 154, 190, 305, 367, 368, 386, 388
Garg, Dinesh, 75, 117, 171, 211
Garimella, Vishal, 64, 107
Garrette, Dan, 370, 389
Gashteovski, Kiril, 155, 194
Gasser, Michael, 616
Gaut, Andrew, 225, 281
Gautam, Akash, 642
Gauthier, Jon, 105, 120, 199, 203
Gawlik, Ireneusz, 74, 115
Ge, Tao, 233, 271
Ge, Yubin, 542, 597
Gebhardt, Kilian, 625
Gebreselassie, Tewodros, 616
Geffen Lan, Nur, 343, 403
Gehrmann, Sebastian, 11, 14, 336, 396
Gelderloos, Lieke, 29, 203
Gella, Spandana, 446, 586, 632
Gemulla, Rainer, 155, 194
Genabith, Josef van, 37, 95, 113, 114, 152, 247, 304, 475, 575, 620
Geng, Ruiying, 69, 111
Geng, Xinwei, 226, 263
Geng, Yuxia, 227, 263
Georgi, Ryan, 616
Gerard, Libby, 638
Gerlach, Andrew, 640
Germann, Ulrich, 619
Gerz, Daniela, 31, 109, 175, 215, 621
Gessler, Luke, 527, 550
Getahun, Dr. Fekade, 616
Geva, Mor, 55, 72, 305, 368
Ghaeini, Reza, 444, 502
Ghannay, Sahar, 632, 633
Ghazvininejad, Marjan, 166, 187, 540, 580
Gholipour Ghalandari, Demian, 79, 137
Ghosal, Deepanway, 233, 271
Ghosh, Debanjan, 541, 597, 627, 638
Ghosh, Kripa, 643
Ghosh, Saptarshi, 643
Giannitsarou, Chryssi, 97, 117
Gienapp, Lukas, 373, 392
Gildea, Daniel, 625
Gilkerson, James, 623
Gimpel, Kevin, 173, 192, 363, 400, 633, 634, 638
Ginter, Filip, 625
Giri, Ritwik, 620
Giryes, Raja, 169, 209
Giulianelli, Mario, 268, 308
Glass, James, 133, 147, 157, 174, 184, 196, 242, 260, 322, 343
Glass, Michael, 171, 211
Glavaš, Goran, 93, 133, 465, 508, 562, 568, 632, 633
Gligoric, Milos, 128, 186
Gliozzo, Alfio, 151, 171, 191, 211
Glover, John, 79, 137
Goharian, Nazli, 137, 178
Golab, Lukasz, 541, 581
Goldberg, Yoav, 40, 46, 59, 64, 100, 139, 180, 198, 283, 305, 322, 337, 368, 460, 485, 505, 526, 535, 543, 563, 622
Goldwasser, Dan, 362, 382, 625
Goldwater, Sharon, 105, 145
Golik, Pavel, 619, 620
Gomes, Christian, 546, 584
Goncalves, Marcos, 546, 584
Gonen, Hila, 46, 64, 485, 505
Gong, Chengyue, 245, 405
Gong, Hongyu, 450, 529, 627
Gong, Ming, 54, 72, 427, 509
Gong, Yeyun, 451, 488

González Ochoa, Simón, 638
González-López, Samuel, 638
Gonzalo, Julio, 267, 371
Goodman, Michael Wayne, 454, 534
Goodman, Noah, 345, 404
Goodwin, Emily, 145, 203
Gordon, Mitchell, 633, 634, 636
Gordon-Hall, Gabriel, 87, 125
Gorinski, Philip John, 87, 125
Gorman, Kyle, 639
Gormley, Matthew R., 549, 604
Gosangi, Rakesh, 642
Gottumukkala, Ananth, 54, 154
Gould, Stephen, 645
Govindarajan, Venkata Subrahmanyan, 155, 194
Goyal, Kartik, 226, 346
Goyal, Naman, 540, 552, 580, 589
Goyal, Pawan, 640, 642
Goyal, Tanya, 34, 166
Goyal, Vikrant, 136
Goyzueta, Aaron, 639
Grünewald, Stefan, 626
Grünigen, Dirk von, 54, 72
Graça, Miguel, 636
Grangier, David, 500, 525, 597, 601
Grave, Edouard, 552, 589
Gray, Jonathan, 341, 384
Grefenstette, Edward, 632
Gregory, Hunter, 628
Grundkiewicz, Roman, 637
Gruzitis, Normunds, 274, 313
Gu, Jiatao, 562, 587, 619
Gu, Nianlong, 455, 534
Gu, Shuhao, 644
Guan, Jian, 540, 580
Guan, Saiping, 430, 511
Guan, Yingjun, 101, 295
Guan, Yong, 54, 72
Gudkov, Vadim, 636
Guerini, Marco, 74, 115
Guestrin, Carlos, 348, 390
Gui, Lin, 251, 291
Guimarães, Patrick, 616
Gulordava, Kristina, 282, 302
Gunasekara, Chulaka, 363, 400
Guo, Daya, 429, 511
Guo, Fengyu, 32, 127
Guo, Jiafeng, 430, 511
Guo, Jiaxian, 620
Guo, Junbo, 46, 64
Guo, Junliang, 37, 113
Guo, Lei, 566, 603
Guo, Mandy, 140, 199
Guo, Mengxue, 620
Guo, Mingming, 642
Guo, Quan, 523, 585
Guo, Shaoru, 54, 72
Guo, Sheng, 482, 499
Guo, Xiaoxiao, 630
Guo, Yufan, 622
Guo, Yuhang, 644
Guo, Zhijiang, 91, 131
Gupta, Abhijit, 540, 597
Gupta, Abhinav, 632, 633
Gupta, Abhirut, 230, 288
Gupta, Ankit, 55, 72, 305, 368
Gupta, Ashim, 640
Gupta, Itika, 641
Gupta, Kshitij, 627
Gupta, Mansi, 343, 403
Gupta, Nitish, 105, 203, 367, 386
Gupta, Shashank, 642
Gupta, Vivek, 155, 194, 633, 634
Gurevych, Iryna, 75, 88, 116, 127, 287, 328, 349, 390, 467, 506, 569, 590
Gururangan, Suchin, 551, 606
Guta, Andreas, 619
Guu, Kelvin, 346, 405
Guy, Ido, 642
Guzmán, Francisco, 75, 116, 435, 531, 552, 589
Gwinnup, Jeremy, 619

H Kumar, shachi, 645
Hätty, Anna, 175, 215
Ha, Le An, 176, 215
Ha, Thanh-Le, 619
Habash, Nizar, 527, 549, 550, 604
Habibi, Amir Ahmad, 639
Hachey, Ben, 423, 483
Haddow, Barry, 307, 350
Haffari, Gholamreza, 228, 264, 425, 465
Hagen, Matthias, 373, 392
Hagiwara, Masato, 566, 602, 620
Hahn, Michael, 40, 99, 243, 322
Hahn, Stefan, 641
Hahn, Udo, 74, 116
Hahn-Powell, Gus, 159, 554
Hahnloser, Richard H.R., 92, 132, 455, 534
Hajishirzi, Hannaneh, 11, 15, 54, 154, 368, 388, 502, 544, 583, 599, 632
Hakkani-Tur, Dilek, 621
Halbe, Shaunak, 645
Haldar, Rajarshi, 565, 602
Hall Maudslay, Rowan, 321, 342, 492, 531, 628
Ham, Donghoon, 47, 124
Hamborg, Felix, 78, 159
Hamill, Chris, 627
Hamilton, Kathleen, 552, 589
Hamilton, William L., 164, 184, 633, 634
Hammerla, Nils, 551, 588
Han, Hou Jeung, 619
Han, Ji Yoon, 625
Han, Jialong, 34, 110, 250, 309
Han, Jiawei, 101, 295, 349, 408, 546, 584
Han, Na-Rae, 625
Han, Qinghong, 423, 501
Han, Song, 525, 601
Han, Wenjuan, 249, 391
Han, Xiaochuang, 366, 403
Han, Xu, 443, 501
Han, Yugui, 32, 127
Hansen, Eric, 619
Hao, Jie, 619
Hao, Yuexing, 35, 129
Haque, Rejwanul, 636
Harbecke, David, 98, 431
Harbusch, Karin, 616
Hardefeldt, Laura, 623
Harris, Kristina, 641
Harte, Naomi, 164, 206
Hartmann, Nathan, 616
Hartvigsen, Thomas, 321, 342
Haruta, Izumi, 312, 530
Hasan, Md Kamrul, 157, 196
Hase, Peter, 366, 385
Hasegawa, Marcello, 566, 603
Hashemi, Jordan, 621
HASHEMPOUR, REYHANEH, 616
Hashimoto, Tatsunori, 50, 186
Hassan Awadallah, Ahmed, 148, 151, 184, 191, 224, 401, 566, 603, 635
Hassan, Hany, 636

Hauer, Bradley, 639
Hauptmann, Alexander, 548, 586
Hautamäki, Ville, 136, 491
Havrylov, Serhii, 625
Hayashi, Hiroaki, 636
Hayashi, Tomoki, 436, 495
Hazan, Tamir, 170, 210
Hazarika, Devamanyu, 233, 271
He, Bolei, 271, 332
He, Chao, 234, 272
He, Daqing, 541, 581
He, Han, 625
He, Hangfeng, 570, 590
He, He, 352, 411
He, Huang, 31, 109
He, Jiacong, 547, 585
He, Jianshan, 30, 86
He, Jun, 460, 501
He, Keqing, 48, 125
He, Liang, 263, 324
He, Pengcheng, 150, 179, 190, 216
He, Ruifang, 32, 127
He, Tianxing, 147, 184
He, Xiaodong, 79, 138, 170, 210
HE, XIAOFENG, 250, 268
He, Xuanli, 228, 264
He, Yulan, 36, 70, 260, 299
He, Zhongjun, 644
Heafield, Kenneth, 93, 134, 307, 350, 525, 562, 636, 637
Hearst, Marti A., 349, 354, 408, 413, 575, 611, 638
Heinecke, Johannes, 625
Helbig, David, 643
Helcl, Jindřich, 636
Hellman, Scott, 638
Hellwig, Oliver, 640
Henderson, James, 435, 569, 590, 609
Henderson, Matthew, 31, 109, 621
Hendrycks, Dan, 171, 211
Hennig, Leonhard, 90, 91, 130, 131, 642
Heo, Yu-Jung, 631
Herbig, Nico, 95, 152, 475, 575
Hermann, Karl Moritz, 632
Herold, Christian, 619
Hershcovich, Daniel, 626
Herzig, Jonathan, 290, 310
Hewitt, John, 367, 385
Hey, Tobias, 289, 329
Hidayat, Albert, 642
Hidey, Christopher, 565, 602
Higy, Bertrand, 282, 301
Hildebrandt, Jordan, 105, 146
Hingerl, Johannes, 75, 116
Hingmire, Swapnil, 629
Hira, Noor-e-, 637
Hirasawa, Tosho, 78, 472, 636
Hirschberg, Julia, 157, 196
Hisamoto, Sorami, 243, 321
Hiware, Kaustubh, 643
Hnatovskiy, Vladislav, 475, 575
Hoang, Hieu, 307, 350
Hockenmaier, Julia, 167, 207, 565, 602, 639
Hofmann, Valentin, 71, 135, 486, 527
Hogan, Julien, 622
Hoi, Steven C.H., 55, 72, 337, 377, 422, 480
Hokamp, Chris, 79, 137
Hong, Yu, 251, 311
Honovich, Or, 502, 598
Hoover, Benjamin, 336, 396
Hope, Tom, 435, 531
Hopkins, Torin, 640
Hoque, Ehsan, 157, 196
horvitz, Eric, 145, 204
Horvitz, Zachary, 627
Hossain, Md Mosharaf, 552, 589
Hossain, Nabil, 166, 187, 376, 414
Hosseini, saghar, 148, 184, 224, 401
Hosseinia, Marjan, 643
Hou, Feng, 460, 501
Hou, Lei, 231, 288, 423, 460
Hou, Xinwen, 464, 504
Hou, Yufang, 88, 127
Hou, Yutai, 86, 108
Hovy, Dirk, 11, 21, 94, 114, 356, 395
Hovy, Eduard, 361, 369, 381, 407, 501, 582
Hsi, Andrew, 354, 413
Hsieh, Cho-Jui, 169, 209, 356, 395, 447, 485
Htut, Phu Mon, 179, 216, 355, 394
Hu, Changjian, 642
Hu, Chi, 447, 485, 637
Hu, Guoping, 42, 86, 108, 120
Hu, Hexiang, 167, 207
Hu, Jennifer, 105, 120, 199, 203
Hu, Jiaying, 440, 480
Hu, Jinglu, 56, 97
Hu, Junjie, 548, 586
Hu, Linmei, 287, 328
Hu, Min, 480, 520
Hu, Pengwei, 479, 519
Hu, Ronghang, 631
Hu, Wenpeng, 546, 584
Hu, Xuemeng, 36, 70
Hu, Yakun, 288, 328
Hu, Yue, 38, 133
Hu, Yuhuang, 92, 132
Hu, Zhifeng, 625
Huang, Biqing, 445, 502
Huang, Chao-Wei, 49, 65, 252, 333
Huang, Chen, 426, 467
Huang, Chu-Ren, 627
Huang, Fei, 237, 276, 540, 543, 580, 582, 619
Huang, Guoping, 37, 132, 563, 601
Huang, Haifeng, 231, 289
Huang, Haoran, 53, 96
Huang, Hen-Hsen, 32, 127
Huang, Jimmy Xiangji, 424, 461
Huang, Jing, 170, 210
Huang, Junhong, 30, 87
Huang, Kaili, 47, 124, 479, 519
Huang, Liang, 38, 174, 192, 193, 644
Huang, Longtao, 233, 272
Huang, Luyang, 353, 412
Huang, Minlie, 47, 48, 124, 125, 255, 266, 295, 305, 479, 519, 540, 580
Huang, Po-Yao, 548, 586
Huang, Qingbao, 484, 523
Huang, Ruihong, 363, 400, 629
Huang, Shujian, 30, 50, 58, 89, 124, 246, 285, 375
Huang, Xiao, 149, 189, 559, 599
Huang, Xiaoxi, 627
Huang, Xin, 230, 288
Huang, Xinting, 49, 125
Huang, Xuanjing, 236, 274, 424, 432, 447, 460, 462, 473, 485, 522
Huang, Yalou, 230, 328
Huang, Yi, 480, 520
Huang, Yichen, 165, 186
Huang, Yiqi, 619
Huang, Yongfeng, 226, 324
Huang, Yungui, 622
Huang, Yuxin, 225, 281
Hubin, Aliaksandr, 90, 130
Hulden, Mans, 639, 640

Humeau, Samuel, 164, 205
Hung, Shyh-Shiun, 32, 127
Hupkes, Dieuwke, 38, 192
Hur, Brian, 623
Hutchinson, Ben, 364, 402
Hwang, Jena D., 625
Hwang, Sung Ju, 34, 110

Ibrahim, Zina, 622
Ie, Eugene, 167, 207
Ifrim, Georgiana, 79, 137
Iida, Shohei, 620
Imankulova, Aizhan, 636
Inaguma, Hirofumi, 436, 495
Indurthi, Sathish Reddy, 619
Inoue, Naoya, 450, 529
Inui, Kentaro, 40, 47, 86, 99, 235, 254, 273, 287, 328, 335, 429, 444, 450, 483, 490, 491, 529
Ippolito, Daphne, 110, 185, 500, 597, 636
Iqbal, Nauman, 617
Ishibashi, Yoichi, 58, 273
Ishwar, Prakash, 566, 603
Isik, Umut, 620
Islam, Mashrukh, 617
Iso, Hayate, 33, 67
Isonuma, Masaru, 51, 70
Iter, Dan, 346, 405
Ito, Takumi, 40, 99
Ittycheriah, Abe, 636
Iurshina, Anastasiia, 632, 633
Ives, Zachary, 299, 382
Iyatomi, Hitoshi, 235, 530
Iyyer, Mohit, 525, 587, 629

Jacobs, Cassandra L., 617
Jacobsen, Nicolai, 157, 196
Jacovi, Alon, 283, 322
Jacovi, Michal, 471, 513
Jagadish, H., 635
Jagannatha, Abhyuday, 149, 189
Jaimes, Alejandro, 629
Jain, Akriti, 627
Jain, Naman, 632, 633
Jain, Rajiv, 36, 111
Jain, Sarthak, 301, 385, 403, 502, 535, 576, 599
Jain, Vihan, 167, 207
Jaiswal, Mimansa, 177, 530
Jaiswal, Nikhil, 627
Jamshid Lou, Paria, 252, 292
Jang, Youngsoo, 47, 124
Jasbi, Masoud, 638
Jastrzebski, Stanislaw, 616
Javdan, Soroush, 627
Jawahar, Ganesh, 46, 64
Jawale, Parth Anand, 552, 606
Jawanpuria, Pratik, 228, 304, 632, 633
Jayannavar, Prashant, 167, 207
Jayarao, Pratik, 639
Jenne, Sabrina, 418, 494
Jensen, David, 362, 382
Jensson, Arnar, 638
Jeon, Hwisang, 250, 269
Jeong, Woo Tae, 34, 110
Jeretic, Paloma, 569, 590
Jernite, Yacine, 341, 384
Ji, Baijun, 92, 132
Ji, Donghong, 290, 310, 447, 464, 469, 509
Ji, Heng, 140, 167, 180, 207, 543, 582, 629
Ji, Jianshu, 179, 216
Ji, Yangfeng, 136, 367, 403, 572

Jia, Chen, 423, 461
Jia, Hao, 92, 132
Jia, Robin, 171, 211, 369, 389
Jia, Shengbin, 643
Jia, Shengyu, 224, 401
Jia, Xin, 429, 490
Jia, Yanan, 642
Jia, Zixia, 452, 510
Jian, Weiyu, 30, 86
Jiang, Chao, 541, 581
Jiang, Daxin, 54, 72, 427, 429, 450, 451, 488, 509, 511, 528
Jiang, Enyi, 101, 295
Jiang, Haoming, 114, 150, 173, 190
Jiang, He, 149, 189
Jiang, Jing, 55, 73, 242, 299
Jiang, Jyun-Yu, 52, 213
Jiang, Lu, 425, 465
Jiang, Meng, 368, 388, 624
Jiang, Menghan, 627
Jiang, Nan, 636
Jiang, Shaohua, 53, 96
Jiang, Songcheng, 26, 179
Jiang, Xin, 34, 68
Jiang, Yong, 236, 237, 274, 276, 626
Jiang, Yufan, 246, 285, 447, 485
Jiang, Zhengbao, 149, 189, 427, 589
Jiang, Zhengping, 571, 591
Jiang, Zhongtao, 444, 461
Jiang, Zhuoren, 230, 287
Jiao, Fangkai, 266, 305
Jimenez Gutierrez, Bernal, 305, 407
Jimeno Yepes, Antonio, 623
Jin, Di, 353, 412, 488, 529, 621
Jin, Hanqi, 433, 452, 469, 474
Jin, Huiming, 449, 605
Jin, Lesheng, 430, 511
Jin, Lifeng, 625
Jin, Shuning, 346, 405
Jin, Shuo, 464, 504
Jin, Tian, 467, 506
Jin, Xiaolong, 430, 511
Jin, Xisen, 364, 401
Jin, Yonghao, 160, 357
Jin, Zhijing, 353, 412
Jindal, Gaurav, 632, 633
Jing, Liping, 230, 288
Jing, Wei, 446, 484
Jo, Jae-young, 243, 302
Jo, Jason, 375, 592
Johnson, Kristen, 617
Johnson, Mark, 252, 292
Johnson, Matthew S., 638
Jojic, Nebojsa, 523, 585
Jones, Erik, 171, 211
Joseph, Kenneth, 299, 382
Joshi, Mandar, 346, 405
Joshi, Pratik, 435, 492
Joshi, Reenam, 622
Joty, Shafiq, 55, 72, 224, 229, 236, 274, 319, 326, 447, 464
JU, Da, 164, 206
JU, QI, 427, 509
Juan, Da-Cheng, 464, 505
Jung, Kerstin, 435, 609
Jung, Kyomin, 52, 95
Jung, Taehee, 170, 210
Jurafsky, Dan, 2, 346, 364, 401, 405
Juzek, Tom S, 620
JV, Kameshwar Rao, 645
Jyothi, Preethi, 252, 333, 620

Köhler, Joachim, 90, 130

Kabiito, David, 616
Kacarevic, Zorica, 418, 494
Kafle, Kushal, 547, 585
Kahn, Andrea, 621
Kaiser, Nicolas, 54, 72
Kajdanowicz, Tomasz, 617
Kajiwara, Tomoyuki, 36, 70
Kallmeyer, Laura, 628
Kallumadi, Surya, 642
Kalra, Kanika, 507, 567
Kamal, Eslam, 495, 555
Kamath S, Sowmya, 645
Kamath, Amita, 369, 389
Kamath, Sanjay, 622
Kameswari, Lalitha, 643
Kamran, Amir, 307, 350
Kan, Min-Yen, 67, 89, 128, 129, 224, 319
Kanan, Christopher, 547, 585
Kanayama, Hiroshi, 51, 69
Kaneko, Masahiro, 287, 328, 636
Kanerva, Jenna, 625
Kang, Daniel, 50, 186
Kang, Dongyeop, 170, 210
Kang, Jaewoo, 54, 154, 250, 269
Kang, Yangyang, 230, 287
Kanjaria, Karina, 622
Kann, Katharina, 355, 394, 449, 605, 616, 625, 639
Kantor, Yoav, 270, 331
Kao, Ben, 282, 322
Kao, Kimberly, 638
Karadzhov, Georgi, 242, 260
Karakanta, Alina, 620
Karamanolakis, Giannis, 559, 599
Karan, Mladen, 465, 562
Karande, Shirish, 507, 567
Karargyris, Alexandros, 622
Karidi, Taelin, 74, 115
Karimi Mahabadi, Rabeeh, 569, 590
Karimi, Sarvnaz, 423, 483
Karipbayeva, Aidana, 639
Karita, Shigeki, 436, 495
Karlsson, Börje, 445, 502
Karnin, Zohar, 172, 212
Kasamatsu, Miho, 624
Kashyap, Satyananda, 622
Kashyap, Sidharth, 637
Kasner, Zdeněk, 636
Kassner, Nora, 531, 609
Kato, Takuma, 254, 273
Katsios, Gregorios, 627
Katsumata, Satoru, 58
Kautz, Henry, 376, 414
Kautz, Jan, 636
Kawada, Yasuhide, 624
Kawahara, Daisuke, 254, 294
Kawahara, Tatsuya, 30, 86
Kazemnejad, Amirhossein, 426, 506
Kearns, Edward, 629
Kehai, Chen, 644
Keith, Katherine, 362, 382
Keller, Frank, 105, 138, 178, 204
Kementchedjhieva, Yova, 75, 116
Kemper, Nathan, 627
Kennedy, Brendan, 364, 401
Kenneth, Neta, 74, 115
Kenter, Tom, 549, 604
Kerianto, Muhamad Danang, 642
Kerz, Elma, 638
Keselj, Vlado, 642
Keung, Phillip, 132, 192
Khabsa, Madian, 151, 191, 624
Khademi, Mahmoud, 484, 524
Khandelwal, Dinesh, 75, 117
Khandelwal, Kartikay, 552, 589
Khanna, Rahul, 555, 593
Khanuja, Simran, 248, 267
Khapra, Mitesh M., 283, 302, 621
Khare, Aparna, 157, 196
Kharitonov, Eugene, 301, 342
Khashabi, Daniel, 371, 409, 512, 570
Khatri, Akshay, 627
Khatri, Jyotsana, 620
Khayrallah, Huda, 637
Khetan, Ashish, 172, 212
Khouja, Jude, 624
Kiela, Douwe, 348, 408
Kim, Donghwan, 34, 110
Kim, Doo Soon, 177, 530, 635
Kim, Gunhee, 627
Kim, Gyuwan, 47, 147
Kim, Hansaem, 625
Kim, Hwichan, 78, 472
Kim, Hyounghun, 345, 404
Kim, Kang-Min, 268, 308
Kim, Kee-Eung, 47, 124
Kim, Kyungmo, 98, 453
Kim, Najoung, 363, 400
Kim, Sangha, 619
Kim, Sungdong, 47, 147
Kim, Yeachan, 268, 308
Kim, Yunsu, 636
Kimmel, Gregory, 150, 190
King Chen, Jennifer, 638
King, Irwin, 55, 69, 72, 111, 337, 377
Kinney, Rodney, 350, 409
Kirchhoff, Katrin, 170, 210, 636, 641
Kirefu, Faheem, 93, 134, 307, 350
Kitada, Shunsuke, 235, 530
Kitaev, Nikita, 434, 573
Kiyono, Shun, 287, 328, 436, 495
Klafka, Josef, 343, 387
Klakow, Dietrich, 482, 499
Klebanov, Beata Beigman, 627
Klein, Ayal, 469, 606
Klein, Dan, 153, 214, 434, 573
Klein, Guillaume, 637
Klein, Stav, 492, 531, 640
Klein, Tassilo, 504, 560
Klie, Jan-Christoph, 467, 506
Klinger, Roman, 643
Knight, Kevin, 563, 601, 619
Knill, Kate, 638
Knoertzer, Manon, 635
Ko, Jeongwoo, 270, 373
Kobayashi, Goro, 235, 491
Kobayashi, Ichiro, 632, 634
Kobayashi, Sosuke, 444, 483
Kochkina, Elena, 467, 565
Kochmar, Ekaterina, 638
Kodner, Jordan, 105, 203, 449, 486
Koehn, Philipp, 307, 350
Koenigstein, Noam, 263, 324
Koh, Pang Wei, 149, 189
Kohita, Ryosuke, 51, 69
Kojima, Noriyuki, 167, 207
Kolachina, Prasanth, 274, 313
Koller, Alexander, 355, 394
Kolluru, Keshav, 423, 460
Kolyada, Nikolay, 471, 513
Komachi, Mamoru, 58, 78, 472, 636
Konam, Sandeep, 641
Kondrak, Grzegorz, 639

Kong, Fang, 442, 459
Kong, Lingpeng, 264, 326
Kong, Sheng-yi, 642
Kong, Xiang, 369, 407, 501, 582
Kong, Xiangnan, 321, 342
Konno, Ryuto, 444, 483
Konopnicki, David, 51, 112
Konstas, Ioannis, 90, 130, 353, 412, 523, 547, 585, 636
Koppel, Moshe, 337, 535
Kordjamshidi, Parisa, 523, 585, 631
Koreeda, Yuta, 234, 272
Korhonen, Anna, 175, 215, 250, 309, 465, 562, 632, 633
Koroleva, Anna, 622
Koski, Chris, 640
Kovaleva, Olga, 622
Krüger, Antonio, 95, 152, 475, 575
Kraljevic, Zeljko, 622
Kramer, Jared, 621
Kratochvíl, Jonáš, 620
Krishna, Amrith, 640
Krishnan, Rishabh, 171, 211
Krishnaswamy, Arvindh, 620
Krone, Jason, 621
Krstovski, Kriste, 565, 602
Kruengkrai, Canasai, 423, 461
Kruiper, Ruben, 90, 130
Krumm, John, 376, 414
Kryscinski, Wojciech, 621
Ku, Lun-Wei, 643
Kudugunta, Sneha, 173, 192
Kudva, Gaurav, 628
Kuhlmann, Marco, 625
Kuhn, Jonas, 89, 110, 625, 639
Kulikov, Ilia, 341, 384
Kulkarni, Mayank, 559, 598
Kulmizev, Artur, 274, 313, 626
Kumar Jena, Amit, 627
Kumar, Amardeep, 627
Kumar, Anand, 645
Kumar, Anuj, 621
Kumar, Ashutosh, 500, 542
Kumar, Dhruv, 541, 581
Kumar, Jay, 51, 111
Kumar, Sachin, 636
Kumar, Sawan, 570, 590
Kumar, Sourav, 136
Kumar, Surender, 642
Kumar, Tarun, 627
Kumar, Vaibhav, 488, 528, 633, 634
Kumari, Surabhi, 627
Kunchukuttan, Anoop, 632, 633
Kundu, Souvik, 55, 73
Kuo, Kevin, 628
Kuribayashi, Tatsuki, 40, 99, 235, 444, 483, 491
Kurita, Keita, 171, 211
Kurma, Bhargav, 507, 567
Kurohashi, Sadao, 254, 294, 335, 514
Kurtz, Robin, 625
Kurzweil, Ray, 140, 199
Kushilevitz, Guy, 460, 543
Kusumawardani, Renny Pradina, 617
Kuwabara, Ryosuke, 227, 325
Kuznetsov, Ilia, 349, 390
Kvapilíková, Ivana, 294, 312
Kwak, Haewoon, 242, 260
Kwiatkowski, Tom, 155, 194, 370, 389
Kwon, Heeyoung, 331, 392
Kwon, Hongseok, 636

La Pietra, Marta, 627
Labaka, Gorka, 294, 312, 492, 531
Laban, Philippe, 354, 413, 575, 611
Ladhak, Faisal, 352, 411
LaFlair, Geoffrey T., 566, 602
Lahav, Dan, 270, 331, 471, 513, 622
Lai, Houtim, 38, 93
Lai, Viet Dac, 629
Lai, Yongkui, 86, 108
Lakumarapu, Nikhil Kumar, 619
Lala, Divesh, 30, 86
Lam, Albert Y.S., 66, 126
Lam, Monica, 31, 148
Lam, Wai, 35, 68, 77, 118
Lan, Man, 236, 313
Lan, Ouyu, 149, 189
Lan, Wuwei, 541, 581
Lan, Yunshi, 55, 73
Lane, William, 449, 486
Lange, Lukas, 75, 116, 467, 506, 632, 633
Langlotz, Curtis, 353, 412
Lansing, Larry, 346, 405
Lapata, Mirella, 138, 178, 307, 354, 408, 413
Lappin, Shalom, 29, 85
Lastras, Luis, 363, 400
Latko, Jan, 157, 196
Lau, Jey Han, 29, 85, 224, 281
Lauly, Stanislas, 620
Law, Jax, 140, 199
Lazaridou, Angeliki, 523, 547
Le, Hung, 422, 480
Le, Phong, 502, 583
Leacock, Claudia, 638
Lebanoff, Logan, 177, 530
Lecouteux, Benjamin, 619
Lee, Aaron, 363, 400
Lee, Beomseok, 619
Lee, Dong Bok, 34, 110
Lee, Dong-Ho, 555, 559, 593, 599
Lee, Dongkyu, 48, 65
Lee, Grandee, 52, 95
Lee, Hankyol, 627
Lee, Hung-yi, 425, 466
Lee, Jaejun, 152, 171, 211, 213
Lee, Jennifer, 544, 598
Lee, Jeong-Gwan, 47, 124
Lee, Ji-Ung, 287, 328
Lee, Jinhyuk, 54, 154, 250, 269
Lee, Jong-Hyeok, 636
Lee, Joon-Young, 547, 585
Lee, Junbum, 643
Lee, Kenton, 369, 388, 551, 588
Lee, Kyusong, 60, 160
Lee, Lillian, 573, 608
Lee, Mark, 628
Lee, Mina, 165, 185
Lee, Nayeon, 624
Lee, Roy Ka-Wei, 266, 305
Lee, Sang-Woo, 47, 147
Lee, SangKeun, 268, 308
Lee, Sangwu, 157, 196
Lee, Seanie, 34, 110
Lee, Seong Per, 642
Lee, Seyeon, 555, 593
Lee, Yoonhyung, 52, 95
Lee, Young-Suk, 128, 165
Lehman, Eric, 301, 385, 622
Lei, Jie, 167, 207, 547, 586
Lei, Kai, 424, 462
Lei, Lizhi, 620
Lei, Tao, 169, 209, 367, 386
Leino, Klas, 342, 386
Lemmens, Jens, 628

Lemon, Oliver, 523, 547
Lenz, Barak, 325, 346
Leo, Marie Stephen, 642
Leong, Chee Wee (Ben), 627, 628
Lepori, Michael, 236, 275
Leung, Cane Wing-Ki, 234, 272
Leung, Ho-fung, 484, 523
Levine, Michelle, 157, 196
Levine, Yoav, 325, 346
Levitan, Sarah Ita, 157, 196
Levy, Omer, 226, 346, 405, 540, 580
Levy, Roger, 105, 120, 199, 203, 371, 409
Lewis, Mike, 352, 411, 540, 580
Lewis, Patrick, 488, 528, 632
Lewis, Will, 619
Lhoneux, Miryam de, 626
Li, Bai, 640
Li, Bei, 246, 285, 637
Li, Belinda, 624
Li, Belinda Z., 551, 588
Li, Binhua, 69, 111
Li, Bryan, 352, 411
Li, Changmao, 370, 407, 628
Li, Chen, 287, 328
Li, Cheng-Te, 46, 107, 643
Li, Chenhui, 39, 431
Li, Chenliang, 34, 110, 248, 307, 447, 464
Li, Chia-Yu, 418, 494
Li, Di, 148, 206
Li, Dianqi, 447, 560
Li, Fei, 160, 357
Li, Guanlin, 37, 132
Li, Haizhou, 52, 95
Li, Han, 443, 483
Li, Hang, 33, 53, 67, 96
Li, Hangyu, 48, 65
Li, Haoran, 79, 138, 427, 588
Li, Hongyu, 624
Li, Huayang, 37, 132
Li, Irene Mengze, 365, 402
Li, Jiahuan, 50, 89
Li, Jierui, 37, 132
Li, Jinchao, 255, 295, 495, 555
Li, Jing, 242, 260
Li, Jingyi, 353, 412
Li, Jiwei, 40, 99, 423, 467, 501, 506
Li, Jiyi, 627
Li, Juanxi, 622
Li, Juanzi, 231, 288, 423, 460
Li, Jun, 236, 274
Li, Junyi Jessy, 128, 186, 361, 381
Li, Kun, 471, 513
Li, Lei, 26, 179, 245, 284
Li, Liunian Harold, 169, 209, 356, 395
Li, Manling, 140, 167, 180, 207
Li, Margaret, 341, 384
Li, MengYuan, 30, 87
Li, Ming, 262, 320
Li, Minqin, 644
Li, Peifeng, 442, 459
Li, Peng, 228, 286, 443, 501
Li, Piji, 50, 69, 110, 111
Li, Ping, 373, 392, 544, 582
Li, Qi, 101, 295
Li, Qimai, 66, 126
Li, Qing, 430, 484, 523, 570
Li, Ru, 54, 72
Li, Ruobing, 48, 124
Li, Shengjie, 446, 484
Li, Shoushan, 251, 331
Li, Shuqun, 627
Li, Steven, 628
Li, Sujian, 432, 473
Li, Tao, 552, 606
Li, Toby Jia-Jun, 357, 396
Li, Wei, 432, 473
Li, Xian, 562, 587, 636
Li, Xiang, 255, 295, 620, 644
Li, Xiang Lisa, 170, 210
Li, Xiao, 98, 453
Li, Xiaochun, 620
Li, Xiaoli, 54, 72, 236, 274, 447, 464
Li, Xiaolin, 169, 209
Li, Xiaolong, 31, 125, 443, 483
Li, Xiaonan, 460, 522
Li, Xiaopu, 619
Li, Xiaoya, 40, 99, 423, 501
Li, Xingyu, 159, 375
Li, Xintong, 563, 601
Li, Xuancai, 644
Li, Xutao, 232, 270
Li, Yaliang, 424, 462
Li, Yang, 547, 585
Li, Yangming, 31, 125, 443, 483
Li, Yanyang, 637
Li, Yanzeng, 245, 284
Li, Ying, 422, 426, 467, 479
Li, Yinqiao, 447, 485, 637
Li, Yitong, 504, 560
Li, Yongbin, 48, 65, 69, 111
Li, Yunyao, 635
Li, Zekang, 440, 519
Li, Zhenghua, 233, 236, 275, 291
Li, Zhenhao, 636
Li, Zhenwen, 432, 473
Li, Zhongli, 450, 488
Li, Zhoujun, 425, 465
Liakata, Maria, 467, 470, 510, 565
Lian, Jianxun, 248, 307
Liang, Chao-Chun, 56, 97
Liang, Chen, 114, 173
Liang, Davis, 170, 210, 636
Liang, Hao, 64, 107
Liang, Junjun, 40, 99
Liang, Paul Pu, 365, 402, 645
Liang, Percy, 149, 165, 171, 185, 189, 211, 345, 369, 389, 404
Liang, Runze, 48, 125
Liang, Weixin, 86, 108
Liang, Yaobo, 54, 72, 450, 528
Liang, Zhihao, 35, 129
Liao, Jianxin, 270, 311
Liao, Yi, 34, 68
Liberman, Mark, 644
Libovický, Jindřich, 636
Lichte, Timm, 628
Lichtenstein, Patricia, 627
Liden, Lars, 495, 555
Liem, David, 101, 295
Lim, Ee-Peng, 266, 305
Lim, Yao Chong, 365, 402
Lin, Angela, 348, 408
Lin, Bill Yuchen, 149, 189, 555, 559, 593, 599
Lin, Chen, 622
Lin, Hongfei, 627
Lin, Hui, 48, 124
Lin, Jimmy, 152, 171, 211, 213, 262, 320, 633, 634
Lin, Kevin, 636
Lin, Qian, 55, 73
Lin, Shou-de, 175, 215
Lin, Simon, 544, 598, 622
Lin, Steven, 641

Lin, Thomas, 641
Lin, Xi Victoria, 337, 364, 377, 401
Lin, Xiexiong, 30, 86
Lin, Yankai, 228, 286, 443, 501
Lin, Ye, 637
Lin, Ying, 140, 180, 543, 582
Lin, Zhaojiang, 252, 333
Lin, Zhouhan, 447, 504
Lin, Zijia, 445, 502
Ling, Qing, 471, 513
Ling, Wang, 264, 326
Linzen, Tal, 155, 194, 236, 275, 321, 355, 366, 385, 394
Lioma, Christina, 490, 569
Lipka, Nedim, 565, 602
Lipton, Zachary C., 343, 403
Lison, Pierre, 90, 130
Litman, Diane, 565, 602, 638
Littman, Michael L., 627
Liu, Alexander H., 425, 466
Liu, Bing, 373, 392, 546, 584
Liu, Chen, 452, 470
Liu, Danni, 620
Liu, Danyang, 248, 307
Liu, Dayiheng, 451, 488
Liu, Fei, 177, 530, 622
Liu, Gongshen, 69, 111
Liu, Hairong, 174, 193
Liu, Han, 66, 86, 108, 126
Liu, Hao, 271, 332
Liu, Haokun, 179, 216, 355, 394
Liu, Hui, 246, 285, 432, 473
Liu, Jerry, 628
Liu, Jiachen, 432, 473, 482, 499
Liu, Jiangming, 307, 408, 459, 521
Liu, Jicong, 53, 96
Liu, Jie, 230, 328, 445, 502
Liu, Jin, 424, 461
Liu, Jingjing, 352, 377, 411, 540, 576, 580
Liu, Jinglin, 33, 128, 253, 333
Liu, Jingzhou, 540, 580
Liu, Kaibo, 38, 174, 192, 193
Liu, Kang, 230, 288, 436, 443, 444, 461, 483, 515
Liu, Lemao, 37, 132, 563, 601
Liu, Ling, 639
Liu, Liting, 230, 328
Liu, Liyuan, 149, 189, 349, 408
Liu, Ming-Yu, 636
Liu, Pengfei, 234, 272, 432, 473
Liu, Qian, 87, 109
Liu, Qiang, 245, 405
Liu, Qiuhui, 37, 113, 114, 247, 304
Liu, Qun, 34, 68, 282, 322, 427, 470
Liu, Renjie, 150, 190
Liu, Shengping, 230, 288, 436, 444, 461, 515
Liu, Shiwan, 444, 461
Liu, Shujie, 87, 108, 246, 252, 285, 292
Liu, Tie-Yan, 33, 128, 253, 263, 324, 333
Liu, Ting, 31, 42, 65, 86, 108, 120, 124–126, 205, 226, 263, 422, 440, 450, 479, 519, 528
Liu, Tingwen, 245, 284
Liu, Tongran, 246, 285, 447, 485
Liu, Wei, 450, 528
Liu, Weijie, 427, 509
Liu, Weili, 101, 295
Liu, Weiwei, 284, 325
Liu, Xianggen, 35, 89
Liu, Xiao, 444, 461
Liu, Xiaodong, 150, 179, 190, 216, 509, 552, 632, 634
Liu, Xiaojiang, 48, 50, 65, 110, 169, 209, 422, 479
Liu, Xiaoqing, 373, 392
Liu, Xiaoyuan, 171, 211
Liu, Xiaozhong, 230, 287
Liu, Xinyu, 171, 211
Liu, Xinyue, 573, 608
Liu, Xuebo, 38, 93
Liu, Yang, 159, 228, 264, 375, 644
Liu, Yijia, 86, 108
Liu, Yinhan, 346, 405, 540, 580
Liu, Yinyin, 33, 185
Liu, Yixian, 249, 391
Liu, Yixin, 636
Liu, Yiyu, 248, 307
Liu, Yuchen, 619
Liu, Zeming, 65, 126
Liu, Zequn, 422, 519
Liu, Zheng, 52, 95
Liu, Zhenghao, 147, 184, 490, 511
Liu, Zhengzhong, 501, 582
Liu, Zhijian, 525, 601
Liu, Zhiyuan, 67, 128, 147, 184, 231, 288, 355, 394, 427, 443, 470, 490, 501, 511, 633, 634
Liu, Zhun, 467, 506
Liu, Zihan, 30, 108, 252, 333, 633, 634
Livescu, Karen, 346, 363, 400, 405, 633, 634
Lo, Kyle, 350, 409, 551, 606
Lockard, Colin, 11, 15, 544, 583
Loftsson, Hrafn, 177, 607
Logan IV, Robert L, 150, 190
Long, Dingkun, 69, 111, 449, 486
Long, Quanyu, 243, 302
Lopez, Adam, 105, 145
Lopez-Lopez, Aurelio, 638
Lotfi, Ehsan, 628
LOU, Jian-Guang, 87, 109, 445, 502
Loukina, Anastassia, 638
Lowe, Ryan, 164, 184
Lu, Chao, 231, 289
Lu, Di, 167, 207
Lu, Jianhua, 573, 608
Lu, Kaiji, 342, 386
Lu, Quan, 233, 272
Lu, Wei, 91, 131, 244, 302
Lu, Yao, 352, 411
Lu, Yi-Ju, 46, 107
Lu, Yu, 619
Lu, Zhiyong, 623
Luan, Yixing, 639
Lukasiewicz, Thomas, 282, 301
Lunt, Bryan, 639
Luo, Fan, 622
Luo, Gan, 231, 288
Luo, Jiebo, 228, 264
Luo, Junyi, 231, 288
Luo, Tianyi, 159, 375
Luo, Weihua, 38, 92, 93, 132, 133
Luo, Ying, 443, 522
Luo, Zhunchen, 288, 328
Luu, Alex, 119, 592
Luu, Anh Tuan, 362, 382
Lux, Florian, 418, 494
Lv, Jiancheng, 451, 488
Lynn, Veronica, 361, 381
Lyu, Boer, 430, 511
Lyu, Michael, 55, 72, 337, 377

Mörbitz, Richard, 625
Müller, Thomas, 290, 310
Ma, Chunping, 69, 111
Ma, Cong, 619
Ma, Hao, 151, 191, 624
Ma, Jun, 559, 599
Ma, Mingbo, 38, 174, 192, 193

Ma, Mingyu Derek, 2
Ma, Nianzu, 373, 392
Ma, Rao, 452, 470
Ma, Ruotian, 424, 462
Ma, Shuai, 246, 285
Ma, Shuming, 246, 285, 425, 465
Ma, Shuo, 480, 520
Ma, Wentao, 86, 108
Ma, Xiyao, 169, 209
Ma, Xuezhe, 501, 582
Ma, Xutai, 619
Ma, Yixiao, 252, 292
Ma, Youmi, 452, 510
Maaløe, Lars, 157, 196
Macháček, Dominik, 620
Macherey, Wolfgang, 425, 465, 620
Madaan, Aman, 128, 186
Maddela, Mounica, 348, 390, 541, 581
Madnani, Nitin, 531, 609, 638
Magdy, Walid, 76, 117
Mager, Manuel, 128, 165, 639
Mahajan, Khyati, 617
Mahata, Debanjan, 642
Mahdy, Mohammady, 268, 308
Mahmood, Asad, 152, 213
Maillard, Jean, 551, 588
Majumder, Bodhisattwa Prasad, 444, 599
Majumder, Navonil, 233, 271
Makarov, Peter, 487, 527, 640
Malagò, Luigi, 632, 633
Malamud, Sophia A., 119, 592
Malhotra, Akanksha, 640
Malhotra, Karan, 628
Malioutov, Igor, 515, 593
Malireddy, Chanakya, 98, 119
Mallik, Arnob, 639
Malmasi, Shervin, 642
Malmaud, Jonathan, 371, 409
Malon, Christopher, 504, 560
Mamidi, Radhika, 627, 643
Mamou, Jonathan, 469, 606
Mandic, Marko, 157, 196
Mani, Anirudh, 641
Maniar, Tirth, 98, 119
Manjunatha, Varun, 36, 111
Manning, Christopher D., 50, 89, 160, 216, 353, 367, 385, 412
Mantravadi, Sahitya, 635
Manzini, Tom, 635
Mao, Chengfeng, 157, 196
Mao, Wenji, 232, 253, 270, 292
Mao, Yuning, 349, 408
Mao, Yuren, 284, 325
Mao, Zhendong, 429, 511
Mao, Zhiming, 242, 260
Mao, Zhuoyuan, 335, 514
Marasović, Ana, 551, 606
Marcus, Mitchell, 449, 486
Mardziel, Piotr, 342, 386
Mari, Alda, 271, 331
Marie, Benjamin, 425, 587
Markert, Katja, 352, 411
Markov, Ilia, 628
Markovitch, Shaul, 460, 543
Maronikolakis, Antonios, 299, 361
Marrese-Taylor, Edison, 645
Marshall, Iain, 101, 180, 622
Martin, Lara, 629
Martin, Louis, 330, 348, 485, 560
Martinez Iraola, David, 623
Martins, André F. T., 573, 608, 639
Maru, Marco, 81, 140
Maruf, Sameen, 425, 465
Marusczyk, Anika, 75, 116
Mascio, Aurelie, 622
Masini, Francesca, 627
Mass, Yosi, 51, 112
Mathias, Sandeep, 638
Mathur, Nitika, 350, 390
Mathur, Prashant, 562, 587, 620
Matsoukas, Spyros, 621
Matsubayashi, Yuichiroh, 254, 335
Matsumaru, Kazuki, 79, 137
Matsumoto, Yuji, 625
Matsumura, Lindsay Clare, 638
Matsuo, Yutaka, 645
Matusov, Evgeny, 619, 620
Maxwelll-Smith, Zara, 638
May, Avner, 169, 209
May, Jonathan, 164, 205
Mayfield, Elijah, 638
Mayhew, Stephen, 637
Maynez, Joshua, 137, 178
Mayuranath, Rahul, 636
Mazuecos, Mauricio, 616, 631
Mazumder, Sahisnu, 373, 392
Mazzola, Matt, 495, 555
McAllester, David, 630
McCaffrey, Daniel F., 638
McCallum, Andrew, 636
McCann, Bryan, 364, 401
McCarley, Scott, 75, 117
McCarthy, Arya, 449, 605, 639
McCarthy, Arya D., 449, 486, 562, 587, 617, 637, 639
McCawley, Michael, 75, 117
McCoy, R. Thomas, 155, 194, 236, 275, 321, 366
McCrae, John Philip, 627
McCurdy, Kate, 105, 145
McDonald, Ryan, 137, 178, 623
McDowell, Bill, 637
McElvain, Gayle, 620
McGrath, Liam, 641
McKeown, Kathleen, 352, 411
McKeown, Kathy, 141
McMullin, Kevin, 640
Meaney, J. A., 159, 607
Medentsiy, Volodymyr, 628
Meek, Christopher, 523, 585
Meftah, Sara, 643
Meghwal, Diksha, 616
Meghwanshi, Mayank, 228, 304
Mehri, Shikib, 49, 147, 148, 205
Mehrotra, Ateev, 641
Mehta, Maitrey, 155, 194, 346, 405
Meinhardt, Eric, 146, 204
Meister, Clara, 464, 560, 639
Mekala, Dheeraj, 36, 188
Meladaki, Kalliopi, 95, 152, 475, 575
Melese, Michael, 616
Meng, Fandong, 35, 89, 228, 264, 440, 519
Meng, Fanyang, 252, 292
Meng, Max, 563, 601
Meng, Rui, 541, 581
Meng, Tao, 224, 243, 251, 302, 311, 401
Meng, Yao, 642
Meng, Yuxian, 40, 99, 423, 501
Meng, Zhao, 177, 572
Mengge, Xue, 245, 284
Mensa, Enrico, 508, 568
Mentch, Lucas, 170, 210
Merck, Derek, 353, 412
Merlo, Paola, 625

Merrill, William, 40, 198
Mersha, Amanuel, 616
Mesgar, Mohsen, 88, 127
Metze, Florian, 633, 634
Metzler, Donald, 35, 186
Meyer, Christian M., 287, 328
Miao, Ning, 245, 284
Miao, Shen-yun, 56, 97
Miaschi, Alessio, 632, 633, 638
Michael, Julian, 469, 606
Michel, Paul, 171, 211
Mielke, Sabrina J., 93, 133
Mieskes, Margot, 11, 17
Mihalcea, Rada, 233, 271
Mihindukulasooriya, Nandana, 151, 191
Miller, Ben, 629
Miller, Timothy, 622, 641
Min, Junghyun, 155, 194
Min, Martin Renqiang, 504, 560
Min, So Yeon, 622
Minaei-Bidgoli, Behrouz, 627
Minervini, Pasquale, 282, 301
Mineshima, Koji, 312, 429, 490, 530
Mirza, Diba, 225, 281
Mirzaee, Roshanak, 631
Mishra, Ajay, 549, 604
Mishra, Bamdev, 228, 304, 632, 633
Mishra, Piyush, 640
Mishra, Pushkar, 288, 328
Mita, Masato, 254, 287, 328, 335
Mitamura, Teruko, 629
Mitchell, Tom, 357, 396
Mitnik, Veronika, 74, 115
Mitrofanova, Olga, 636
Mittal, Arpit, 624
Miyawaki, Shumpei, 254, 273
Mizumoto, Tomoya, 254, 335
Mizzaro, Stefano, 267, 371
Mlynchyk, Katsiaryna, 54, 72
Moghe, Nikita, 621
Mohamed, Abdelrahman, 540, 580
Mohammad, Saif M., 355, 376, 394, 414, 533, 609
Mohammadi, Pouya, 628
Mohananey, Anhad, 625
Mohankumar, Akash Kumar, 283, 302
Mohiuddin, Tasnim, 229, 326
Monroe, Will, 637
Monshizadeh, Mahsa, 95, 152, 475, 575
Monz, Christof, 636
Moon, Jihyung, 643
Mooney, Raymond, 128, 186
Moosavi, Nafise Sadat, 569, 590
Moradshahi, Mehrad, 31, 148
Morariu, Vlad, 36, 111
Morency, Louis-Philippe, 157, 196, 365, 402, 467, 484, 506, 586, 645
Moreno, Ryan, 559, 599
Morgan, Jonathan, 299, 382
Mori, Junichiro, 51, 70
MORICEAU, Véronique, 271, 331
Morio, Gaku, 234, 272
Morishita, Terufumi, 234, 272
Moschitti, Alessandro, 369, 407
Moslem, Yasmin, 636
Mostafazadeh Davani, Aida, 364, 401
Mou, Lili, 11, 18, 35, 89, 352, 411, 541, 581
Mou, Xiangyang, 630
Movshovitz-Attias, Dana, 270, 373
Mozaffari, Sahand, 639
Mroczkowski, Robert, 74, 115
Mroueh, Youssef, 262, 320
Mu, Jesse, 345, 404
Muchovej, John, 177, 530
Mueller, Aaron, 366, 385
Mueller, David, 544, 599
Mueller, Erik, 621
Mukherjee, Animesh, 64, 107
Mukherjee, Arjun, 643
Mukherjee, Subhabrata, 151, 191, 224, 401, 566, 603
Mukherjee, Sudipto, 566, 603
Mukherjee, Vandana Mukherjee, 622
Muller, Benjamin, 74, 115, 485, 560
Mulugeta, Wondwossen, 616
Mulyar, Andriy, 565, 602
Munawwar, Eram, 642
Muradoglu, Saliha, 235, 273, 640
Murawaki, Yugo, 254, 294
Muresan, Smaranda, 541, 565, 597, 602, 617, 627
Murikinati, Nikitha, 639, 640
Murray, Kenton, 637
Murray, William, 638
Murty, Shikhar, 149, 189
Murugan, Srikala, 551, 606
Myaeng, Sung-Hyon, 243, 302
Myers, Brad, 357, 396

N T V, Satya Dev, 632, 633
Névéol, Aurélie, 11, 17
Nabi, Moin, 504, 560
Nachman, Lama, 645
Nadkarni, Prajit, 642
Nagata, Masaaki, 620
Nagesh, Ajay, 563, 601, 619
Nagoudi, El Moatez Billah, 636
Naidu, Suresh, 362, 382
Naik, Aakanksha, 522, 583
Najork, Marc, 444, 599
Nakamura, Satoshi, 58, 248, 273, 330, 619, 620
Nakashole, Ndapa, 248, 348
Nakatumba Nabende, Joyce, 616
Nakayama, Hideki, 227, 325, 442, 459
Nakov, Preslav, 46, 64, 242, 248, 260, 330, 418, 515
Nallani, Sneha, 136, 514
Nallapati, Ramesh, 306, 369
Nallasamy, Udhyakumar, 526, 563
Namysl, Marcin, 90, 130
Nan, Guoshun, 91, 131
Napierski, Daniel, 140, 180
Narasimhan, Sharan, 283, 302
Narayan, Shashi, 128, 137, 166, 178
Narayan-Chen, Anjali, 167, 207
Narayanan, Shrikanth, 253, 410, 629
Narayanaswamy, Balakrishnan, 148, 184
Naseem, Tahira, 128, 165
Naskar, Subhajit, 636
Nasr, Mohamed, 75, 117
Nasukawa, Tetsuya, 51, 69
Natarajan, Ganapathy, 642
Naudin, Louise, 263, 324
Navigli, Roberto, 81, 140, 175, 215, 330, 349
Nazir, Ambreen, 64, 107
Negri, Matteo, 465, 562, 619, 620
Nejadgholi, Isar, 623
Nelson, Max, 640
Nema, Preksha, 283, 302
Nemade, Gaurav, 270, 373
Nematzadeh, Aida, 167, 207
Nenkova, Ani, 101, 180
Neubig, Graham, 128, 149, 171, 186, 189, 211, 343, 403, 427, 501, 544, 545, 552, 562, 566, 589, 599, 601, 603, 636, 640
Neumann, Guenter, 621, 623

Neumann, Mark, 350, 409
Neumann, Michael, 418, 494, 641
Neves, Leonardo, 555, 593
Newheiser, Anna, 627
Ney, Hermann, 265, 304, 619, 636
Ng, Hwee Tou, 55, 73
Ng, Patrick, 306, 369
Ng, Wilfred, 250, 309
Ngiam, Jiquan, 525, 587
Nguyen, Dai Quoc, 245, 284
Nguyen, Dong, 470, 510
Nguyen, Ha, 619
Nguyen, Hoang, 635
Nguyen, Thai Son, 619
Nguyen, Thai-Son, 620
Nguyen, Thanh, 463, 584
Nguyen, Thanh-Tung, 236, 274, 447, 464
Nguyen, Thien Hai, 423, 461
Nguyen, Thien Huu, 543, 598, 621, 629
Nguyen, Toan Q., 170, 210, 636
Nguyen, Tu, 245, 284
Nguyen, Tuan-Nam, 619
Nguyen, Xuan-Phi, 236, 274, 447, 464
Nicolai, Garrett, 366, 385, 639, 640
Nie, Allen, 351, 375
NIE, Jian-Yun, 566, 603
Nie, Pengyu, 128, 186
Nie, Yixin, 348, 408, 624
Niehues, Jan, 619, 620
Nijkamp, Erik, 249, 391
Nikolaev, Dmitry, 74, 115
Nikolaev, Vitaly, 370, 389
Nikolov, Nikola I., 92, 132, 455, 534
Ning, Qiang, 512, 570, 590
Nishida, Noriki, 442, 459
Niu, Cheng, 440, 479, 482, 499, 519
Niu, Xing, 562, 587
Niu, Yilin, 266, 305
Niu, Zheng-Yu, 65, 124, 126, 205
Nivre, Joakim, 274, 313, 626
Nkolele, Risuna, 617
Nogueira dos Santos, Cícero, 262, 320
Noji, Hiroshi, 243, 282
Nokhiz, Pegah, 155, 194
Nomoto, Tadashi, 636
Norouzi, Mohammad, 228, 264
Nouri, Elnaz, 348, 408
Nowak, Pawel Krzysztof, 290, 310
Nye, Benjamin, 101, 180, 622

O'Connor, Alexander, 499, 541
O'Connor, Brendan, 362, 382
O'Donnell, Timothy J., 145, 203, 447, 504
O'Hara, Nathan, 628
O'Reilly, Randall, 375, 592
Oard, Douglas, 36, 111, 546, 584
Oda, Yusuke, 636
Oepen, Stephan, 625
Ogunbona, Philip O., 233, 271
Oguz, Barlas, 488, 528
Oh, Alice, 441, 520
Oh, Tae Hwan, 625
Ohashi, Sora, 36, 70
Okabe, Shu, 75, 116
Okazaki, Naoaki, 79, 92, 93, 113, 133, 137
Okur, Eda, 645
Oliver, Antoni, 628
Oluwatobi, Olabiyi, 621
Omelianchuk, Kostiantyn, 638
On, Kyoung-Woon, 631
Oncevay, Arturo, 152, 213, 616
Ong, Donovan, 362, 382
Onysko, Alexander, 628
Oprea, Silviu, 76, 117
Orbach, Matan, 471, 513
Ordonez, Vicente, 364, 401
Ore, Brian, 619
Origgi, Gloria, 271, 331
Orii, Lisa, 353, 412
Ortega, Daniel, 418, 494
Ortiz Rojas, Sergio, 307, 350
Ortiz Suárez, Pedro Javier, 74, 97, 115, 485, 560
Otegi, Arantxa, 488, 528
Ott, Myle, 173, 192, 552, 589
Ou, Zhijian, 48, 109
Ouchi, Hiroki, 47, 86, 254, 273, 444, 483
Ouyang, Yawen, 30, 124
Ozaki, Hiroaki, 234, 272

P, Pranav, 627
Padó, Sebastian, 299, 381
Padhi, Inkit, 262, 320
Padnos, Dan, 325, 346
Paetzold, Gustavo Henrique, 616
Pai, Nithish, 642
Pai, Sharan, 39, 273
Pakhomov, Serguei, 145, 203
Pal, Chris, 155, 194, 362, 382
Pal, Santanu, 95, 152, 475, 575
Palaskar, Shruti, 641
Paliwal, Sudarshan, 631
Palmer, Alexis, 552, 589
Palmer, Martha, 552, 606
Palomaki, Jennimaria, 370, 389
Palshikar, Girish, 629
Pan, Li, 230, 287
Pan, Liangming, 67, 89, 128, 129
Pan, Lin, 75, 117, 171, 211
Pan, Xiaoman, 140, 180
Pan, Xingyuan, 346, 405
Pan, Yirong, 98, 453
Pan, Yue, 641
Pandramish, Vinay, 351, 592
Pang, Bo, 249, 391
Pang, Richard Yuanzhe, 173, 192, 355, 394
Pannier, Baptiste, 263, 324
Pant, Kartikey, 627
Panthaplackel, Sheena, 128, 186
Papagelis, Manos, 374, 393
Papalampidi, Pinelopi, 138, 178
Papangelis, Alexandros, 621
Pappas, Dimitris, 623
Paraskevopoulos, Georgios, 157, 196
Parekh, Tanmay, 128, 186
Parekh, Zarana, 525, 587
Parikh, Ankur, 540, 580
Parikh, Soham, 636
Paris, Cecile, 423, 483
Park, Hyeryun, 98, 453
Park, Junsu, 636
Park, Seokwon, 625
Park, Seongkeun, 98, 453
Park, Thomas, 495, 555
Park, Yangsook, 621
Paroubek, Patrick, 622
Parra, Denis, 616
Parthasarathi, Prasanna, 164, 184
Parthasarathy, Srinivas, 157, 196
Pasini, Tommaso, 269, 308
Patil, Sangameshwar, 629
Patra, Aditya, 293, 334
Patro, Jasabanta, 64, 107

Patro, Sohan, 643
Patwardhan, Manasi, 507, 567
Patwary, Mostofa, 30, 205
Patzer, Rachel E., 622
Paul, Michael J., 151, 191
Paulik, Matthias, 492, 526, 532, 563
Paullada, Amandalynne, 622
Pavlick, Ellie, 11, 14, 155, 194
Pavlopoulos, John, 289, 329
Pawlicka Maule, Anna Paula, 617
Pecina, Pavel, 463, 546
Peddagangireddy, Vishal, 622
Pei, Jiaxin, 34, 110, 248, 307
Peinelt, Nicole, 470, 510
Pejhan, Elham, 626
Pellicanò, Nicola, 263, 324
Pelsmaeker, Tom, 485, 504
Pendus, Cezar, 75, 117
Peng, Baolin, 255, 295, 440, 480, 495, 555
Peng, Hao, 447, 560
Peng, Minlong, 424, 462
Peng, Nanyun, 2, 541, 597, 629
Peng, Tao, 627
Peng, Yifan, 623
Peng, Yihui, 449, 605
Penn, Gerald, 633, 634
Pennebaker, James, 145, 204
Perarnau, Guim, 515, 593
Percha, Bethany, 622
Pereira, Antônio, 546, 584
Pereira, Lis, 632, 634
Peris, Charith, 621
Perl, Tal, 169, 209
Peskov, Denis, 260, 300
Peters, Ben, 639
Petroni, Fabio, 632
Petrou-Zeniou, Panayiota, 366, 385
Petruck, Miriam R L, 435, 531
Peyrard, Maxime, 93, 133
Pham, Nghia The, 79, 137
Pham, Ngoc-Quan, 619
Phan, Minh Hieu, 233, 271
Phang, Jason, 179, 216, 355, 394
Phung, Dinh, 245, 284
Piccinno, Francesco, 290, 310
Pierrehumbert, Janet, 71, 135, 486, 527
Pilan, Ildiko, 638
Pilehvar, Mohammad Taher, 97, 117
Pillay, Siddhanth, 639
Pimentel, Tiago, 307, 321, 342, 371, 449, 486, 492, 527, 531, 550, 628
Pineau, Joelle, 164, 184
Ping, Peipei, 101, 295
Pino, Juan, 619
Pinter, Yuval, 301, 403
Pinto, Joel, 641
Pitler, Emily, 155, 194
Pitrelli, John, 75, 117
Pla Sempere, Leopoldo, 307, 350
Poczos, Barnabas, 128, 186
Poerner, Nina, 469, 510
Poesio, Massimo, 444, 483
Polák, Peter, 620
Poliak, Adam, 571, 591
Polozov, Oleksandr, 509, 523, 552, 585
Polymenakos, Lazaros, 621
Poncelas, Alberto, 265, 326
Poon, Hoifung, 179, 216
Poria, Soujanya, 233, 271, 645
Portelli, Beatrice, 624
Post, Matt, 243, 321, 637
Potapczyk, Tomasz, 619
Potapenko, Anna, 523, 547
Potdar, Saloni, 630
Potthast, Martin, 74, 115, 291, 311, 373, 392
Potti, Navneet, 444, 599
Pouran Ben Veyseh, Amir, 543, 598, 621
Prabhakaran, Vinodkumar, 364, 402
Prabhu, Nikhil, 639
Prabhumoye, Shrimai, 128, 171, 186, 211
Prada, Jonathan, 455, 534
Pranav A, 486, 527
Prasad, Archiki, 252, 333
Preotiuc-Pietro, Daniel, 299, 361, 522, 559, 583, 598
Press, Ofir, 226, 405
Provilkov, Ivan, 132, 173
Provost, Emily Mower, 177, 530
Prud'hommeaux, Emily, 638
Pruksachatkun, Yada, 179, 216, 355, 394
Pruthi, Danish, 343, 403
Przybysz, Pawel, 619
Pujar, Saurabh, 75, 117
Puri, Raul, 30, 205
Purver, Matthew, 29, 85
Pyysalo, Sampo, 625

Qi, Fanchao, 427, 470
Qi, Jianzhong, 49, 125
Qi, Peng, 160, 216
Qi, Qi, 270, 311
Qi, Tao, 226, 248, 307, 324
Qi, Yuan, 53, 96
Qian, Jing, 225, 281
Qian, Kun, 635
Qian, Peng, 105, 120, 199, 203
Qian, Tieyun, 251, 291
Qian, Yusu, 39, 592
Qiao, Chao, 33, 67
Qiao, Ying, 248, 307
Qiao, Yu, 638
Qin, Bing, 226, 263
Qin, Lianhui, 364, 401
Qin, Libo, 31, 125, 440, 519
Qin, Qi, 546, 584
QIN, Tao, 253, 333
Qin, Ying, 620
Qiu, Shuang, 450, 528
Qiu, Xipeng, 236, 274, 432, 460, 473, 499, 522, 580
Quan, Jun, 480, 520
Quan, Xiaojun, 233, 311, 450, 471, 480, 513, 520, 528, 549, 604
Quirk, Chris, 635

Rücker, Susanna, 74, 116
Ré, Christopher, 169, 209, 464, 505
Radev, Dragomir, 540, 597
Radicioni, Daniele P., 508, 568
Rae, Jack, 504, 560
Rafferty, Anna, 633, 634
Rafieian, Bardia, 312, 572
Raganato, Alessandro, 264, 326
Raghavan, Preethi, 622
Raghunathan, Aditi, 171, 211
Rahman, Wasifur, 157, 196
Raina, Vatsal, 638
Rajaby Faghihi, Hossein, 631
Rajamanickam, Santhosh, 288, 328
Rajani, Nazneen Fatema, 301, 364, 385, 401, 540, 597
Rajda, Krzysztof, 617
Ram, Achyudh, 633, 634
Ram, Ori, 325, 346

Ramírez-Sánchez, Gema, 307, 350
Ramamurthy, Ranjani, 641
Ramanarayanan, Vikram, 641
Rambow, Owen, 532, 609
Ramesh, Krithika, 617
Rameshkumar, Revanth, 353, 412
Ramnath, Rajiv, 622
Ramrakhiyani, Nitin, 629
Ramtej, Jaidam, 642
Ran, Qiu, 228, 286
Ranta, Aarne, 274, 313
Ranzato, Marc'Aurelio, 173, 192
Rao, Nikhil, 463, 584
Rao, Sudha, 348, 408
Rao, Yanghui, 430, 570
Rao, Yuan, 64, 107
Rastogi, Abhinav, 621
Rathore, Vipul, 423, 460
Rauber, Andreas, 616
Raunak, Vikas, 633, 634
Ravfogel, Shauli, 485, 505
Ravi, Selvan Sunitha, 645
Ravi, Sujith, 270, 373, 464, 505
Ravindran, Balaraman, 283, 302, 621
Ravishankar, Vinit, 274, 313, 512, 570
Rawat, Bhanu Pratap Singh, 622
Rawlins, Kyle, 544, 582
Ray Chowdhury, Jishnu, 351, 514
Ray, Baishakhi, 352, 411
Razavi, Ali, 504, 560
Reddy, Siva, 446, 586
Rehbein, Ines, 275, 314
Rei, Marek, 175, 176, 215, 632
Reichart, Roi, 170, 175, 210, 215, 573, 608
Reinecke, Katharina, 629
Rejwan, Idan, 263, 324
Ren, Shuo, 246, 285
Ren, Xiang, 149, 189, 349, 364, 401, 408, 555, 559, 593, 599, 629
Ren, Yi, 33, 128, 253, 333, 620
Resnick, Cinjon, 632, 633
Resnik, Philip, 36, 111, 546, 584
Reznik, Ilya, 105, 146
Ri, Ryokan, 56, 97
Ribeiro, Marco Tulio, 348, 390
Richardson, Matthew, 509, 552
Richburg, Aquia, 617
Riedel, Sebastian, 488, 528, 552, 589
Rieser, Verena, 353, 412, 547, 585
Rijhwani, Shruti, 501, 522, 544, 545, 583, 599
Riley, Parker, 525, 601
Riloff, Ellen, 351, 607, 627
Rimell, Laura, 632
Rinaldi, Alex, 29, 85
Rinott, Ruty, 488, 528
Riordan, Brian, 638
Rios, Anthony, 622
Rishe, Naphtali, 629
Ritter, Alan, 33, 67, 348, 361, 381, 390
Rivas Rojas, Kervy, 152, 213
Roark, Brian, 527, 550
Roberts, Angus, 622
Roberts, Kirk, 641
Rocha, Leonardo, 546, 584
Roddy, Matthew, 164, 206
Rodrigo, Alvaro, 54, 72
Rodriguez, Cristian, 645
Roesler, Oliver, 645
Roessler, Oliver, 641
Rohanian, Omid, 176, 215
Roit, Paul, 469, 606
Roitman, Haggai, 51, 112
Rojecki, Andrew, 630
Rokhlenko, Oleg, 642
Roller, Stephen, 164, 206, 341, 384
Romary, Laurent, 97, 115, 485, 560
Ronanki, Srikanth, 641
Rooshenas, Amirmohammad, 636
Rose, Carolyn, 522, 583
Rosenstein, Mark, 638
Rosset, Sophie, 632, 633
Roth, Dan, 11, 20, 299, 371, 382, 409, 512, 570, 590
Rothe, Sascha, 128, 166, 636
Rotman, Guy, 573, 608
Roukos, Salim, 75, 117, 128, 165
Roy, Abhinaba, 233, 271
Roy, Deb, 629
Roy, Kalyani, 642
Rozovskaya, Alla, 638
Ruan, Weitong, 573, 608
Rubin, Alexander, 628
Rubini, Luca, 641
Rubino, Raphael, 425, 587, 636
Ruder, Sebastian, 321, 367, 492, 531
Rudin, Cynthia, 628
Rudzicz, Frank, 640
Ruiz, Miguel, 641
Rumshisky, Anna, 622
Rundensteiner, Elke, 321, 342
Runyon, Christopher, 638
Ruprecht, Thomas, 625
Rush, Alexander, 167, 170, 207, 210, 495, 611
Russin, Jacob, 375, 592
Rust, Phillip, 75, 116
Rust, Steve, 622
Ryan, James, 621
Ryan, Zach, 640
Rybak, Piotr, 74, 115
Ryskina, Maria, 549, 604

S R, Dhanush, 292, 333
Sánchez Villegas, Danae, 299, 361
Søgaard, Anders, 512, 570, 625
Sa, Ning, 627
Sabharwal, Ashish, 371, 409
Sachan, Mrinmaya, 169, 209
Sachdeva, Nikhil, 39, 273
Sachdeva, Prince, 39, 273
Sadat, Fatiha, 643
Sadde, Shoval, 59, 139, 622
Sadeqi Azer, Erfan, 371, 409
Sadeque, Farig, 622
Saeboe, Lilja Maria, 74, 115
Sagae, Kenji, 625, 638
Sagar, Sangeet, 620
Sagot, Benoît, 74, 97, 115, 330, 348, 485, 560
Saha, Sriparna, 293, 334, 443, 460
Saha, Tulika, 293, 334
Sahay, Saurav, 645
Sahu, Sunil Kumar, 268, 308
Saini, Nikhil, 620
Sajed, Tanvir, 376, 414
Sajjad, Hassan, 133, 174, 322, 343
Sakaguchi, Keisuke, 571, 591
Sakata, Ichiro, 51, 70
Sakrajda, Andrzej, 75, 117
Sala, Frederic, 464, 505
Salakhutdinov, Ruslan, 128, 171, 186, 211, 365, 402
Salazar, Julian, 170, 210, 636
Saldias, Belen, 629
Saleh, Abdelrhman, 621
Saleh, Shadi, 463, 546

Salehi, Mohammadreza, 426, 506
Salesky, Elizabeth, 158, 197, 307, 371, 464, 560, 619, 639
Salunke, Devika, 641
Sanchez-Gutierrez, Claudia Helena, 638
Sang, Lan, 640
Santhanam, Sashank, 617
Santus, Enrico, 624
Santy, Sebastin, 435, 492
Sanyal, Soumya, 366, 385
Sap, Maarten, 11, 20, 145, 204, 364, 401, 629
Sarawagi, Sunita, 632, 633
Sardana, Ashish, 645
Sarrías, Elsa, 307, 350
Sarrouti, Mourad, 631
Sartiano, Daniele, 626
Sartran, Laurent, 264, 326
Sasaki, Shota, 254, 335
Sasano, Ryohei, 250, 309
Sato, Shiki, 47, 86
Saunders, Danielle, 525, 526, 563
Savary, Agata, 13
Savkov, Aleksandar, 551, 588
Savoldi, Beatrice, 465, 562
Savova, Guergana, 622
Sawaf, Hassan, 620
Saxena, Apoorv, 305, 368
Scarlini, Bianca, 269, 308
Scarton, Carolina, 330, 348
Schütze, Hinrich, 71, 135, 268, 308, 469, 486, 510, 527, 531, 609
Schüz, Simeon, 446, 585
Schaaf, Thomas, 170, 210
Schallhart, Christian, 549, 604
Scherbakov, Andreas, 640
Schick, Timo, 268, 308
Schijndel, Marten van, 145, 204
Schläpfer, Philippe, 54, 72
Schlechtweg, Dominik, 175, 215
Schluter, Natalie, 5
Schmidt, Maximilian, 418, 494
Schneider, Felix, 619, 620
Schneider, Nathan, 355, 394, 527, 550
Schoenick, Carissa, 535, 576
Schofield, Alexandra, 11, 21
Schröder, Fynn, 226, 284
Schuler, William, 625
Schulte im Walde, Sabine, 175, 215, 616
Schulte, Henri, 624
Schumacher, Elliot, 565, 602
Schumann, Raphael, 352, 411
Schuster, Sebastian, 363, 400
Schuster, Tal, 624
Schwartz, H. Andrew, 356, 361, 381, 395
Schwartz, Roy, 40, 198, 447, 448, 560, 561
Schwenk, Holger, 488, 528
Scontras, Greg, 146, 204
Scozzafava, Federico, 81, 140, 269, 308
Sdun, Regitze, 157, 196
Searle, Thomas, 622
Seddah, Djamé, 46, 64, 74, 115, 485, 560, 625
Seker, Amit, 492, 531
Sekulic, Ivan, 91, 131
Sellam, Thibault, 540, 580
Selo, Vittorio, 515, 593
Semmar, Nasredine, 643
Sen, Cansu, 321, 342
Senellart, Jean, 92, 113, 637
Sengupta, Neha, 268, 308
Sennrich, Rico, 92, 113, 247, 304, 525, 562
Seo, Minjoon, 54, 154
Seppi, Kevin, 105, 146, 628
Serra, Giuseppe, 624
Setiawan, Hendra, 526, 563
Setlur, Amrith, 128, 186
Settles, Burr, 566, 602, 637
Severyn, Aliaksei, 128, 166
Seyoum, Binyam Ephrem, 616
Sha, Fei, 167, 207
Shaar, Shaden, 248, 330, 418, 515
ShafieiBavani, Elaheh, 623
Shafiq, Zubair, 152, 213
Shafran, Izhak, 641
Shah, Deven Santosh, 356, 395
Shah, Rajiv Ratn, 39, 273, 642
Shah, Rushin, 621
Shah, Smit, 642
Shahaf, Dafna, 435, 531
Shahbazi, Hamed, 444, 502
Shahid, Usman, 630
Shahidi, Hamidreza, 262, 320
Shaikh, Samira, 616, 617
Shalev-Shwartz, Shai, 325, 346
Shan, Yong, 440, 519
Shang, Chao, 151, 191
Shang, Hengchao, 620
Shang, Jingbo, 36, 188, 546, 584
Shangipour ataei, Taha, 627
Shani, Chen, 435, 531
Shao, Bin, 263, 324
Shao, Jie, 266, 305
Shao, Junming, 51, 111
Shao, Liqun, 635
Shao, Yutong, 248, 348
Sharaf, Amr, 149, 189, 636
Sharir, Or, 325, 346
Sharma, Arpit, 642
Sharma, Dipti, 136, 514
Sharma, Dipti Misra, 136, 351, 592, 617
Sharma, Piyush, 446, 484
Sharma, Yashvardhan, 627
Shashua, Amnon, 325, 346
Shaw, Peter, 551, 588
Shayandeh, Shahin, 495, 555
Shen, Dinghan, 51, 111, 165, 186, 504, 560
Shen, Jiaming, 546, 584
Shen, Lei, 47, 65
Shen, Ming, 559, 599
Shen, Weizhou, 233, 311
Shen, Xiaoyu, 479, 482, 499, 519
Shen, Yelong, 167, 207, 450, 529
Shen, Yikang, 447, 504
Shen, Ying, 424, 462, 645
Sheng, Qiang, 46, 64
Shenoy, Aman, 645
Sheriff, Michael, 629
Shi, Bowen, 633, 634
Shi, Chuan, 287, 328
Shi, Haoyue, 633, 634
Shi, Shuming, 37, 50, 110, 132, 228, 264
Shi, Tianze, 573, 608
Shi, Tingxun, 619
Shi, Wenxuan, 230, 328
Shi, Xing, 619
Shi, Zhan, 499, 580
Shieber, Stuart, 621, 638
Shillingford, Brendan, 282, 301
Shin, Hejin, 622
Shin, Joongbo, 52, 95
Shin, Richard, 509, 552
Shing, Han-Chin, 546, 584
Shiralkar, Prashant, 11, 15, 544, 559, 583, 599
Shirani, Amirreza, 565, 602

Shivade, Chaitanya, 622, 641
Shlain, Micah, 59, 139, 622
Shmidman, Avi, 337, 535
Shmidman, Shaltiel, 337, 535
Shoeybi, Mohammad, 30, 205
Shoham, Yoav, 325, 346
Short, Tyler, 640
Shou, Lidan, 424, 461
Shou, Linjun, 54, 72, 427, 509
Shoukat, Arslan, 637
Shreevastava, Sagarika, 640
Shrestha, Robik, 547, 585
Shrivastava, Manish, 98, 119, 136, 514, 632, 633
Shterionov, Dimitar, 265, 326
Shu, Chang, 29, 85
Shui, Ruihao, 67, 128
Shukla, Swadheen, 495, 555
Shum, Heung-Yeung, 148, 206
Shum, Michael, 621
Shuster, Kurt, 147, 164, 205, 206
Shutova, Ekaterina, 288, 328, 551, 588, 627, 628
Shwartz, Vered, 11, 20
Si, Luo, 35, 68, 251, 331
Sia, Suzanna, 238, 315
Siddhant, Aditya, 173, 192
Siddiqi, Abdul Basit, 617
Sigdel, Dibakar, 101, 295
Sil, Avi, 75, 117, 171, 211
Silfverberg, Miikka, 639
Simi, Maria, 626
Siminyu, Kathleen, 616
Simonsen, Jakob Grue, 490, 569
Singer, Assaf, 639
Singh, Avinash Kumar, 629
Singh, Gagandeep, 641
Singh, Rishubh, 294, 312
Singh, Sameer, 54, 150, 154, 190, 348, 367, 368, 386, 388, 390
Singla, Karan, 253, 410
Sinha, Aman, 627
Sinha, Koustuv, 145, 164, 184, 203
Sinha, Moumita, 635
Sinha, Sayan, 643
Sirrianni, Joseph, 373, 392
Sitaram, Sunayana, 248, 267
Skjonsberg, Sam, 535, 576
Skurzhanskyi, Oleksandr, 638
Slonim, Noam, 270, 331, 471, 513
Small, Kevin, 33, 67
Smith, Eric Michael, 147, 205
Smith, Noah A., 40, 145, 198, 204, 226, 364, 401, 405, 447, 448, 551, 560, 561, 606, 629
Smith-Stvan, Laurel, 642
Smus, Boris, 376, 414
Snajder, Jan, 632, 633
Snell, Quinn, 105, 146
Sobrevilla Cabezudo, Marco Antonio, 152, 213
Socher, Richard, 55, 72, 224, 301, 319, 337, 377, 385, 540, 597, 621
Soldaini, Luca, 369, 407
Soleymani Baghshah, Mahdieh, 426, 506
Solorio, Thamar, 543, 565, 598, 602
Somasundaran, Swapna, 629
Sommerauer, Pia, 119, 572
Song, Dawn, 171, 211
Song, Haiyue, 335, 514
Song, Haoyu, 422, 479
Song, Linfeng, 363, 400, 542, 597
Song, Ruihua, 566, 603
Song, Sen, 35, 89
Song, Xiaodan, 150, 190
Song, Yan, 471, 513, 549, 604, 641
Song, Yangqiu, 371, 409
Song, Yi, 638
SONG, YIPING, 48, 65, 422, 519
Song, Yonghao, 440, 479
Song, Yuxuan, 245, 284
Sorensen, Jeffrey, 289, 329
Soricut, Radu, 33, 185, 446, 484
Soroa, Aitor, 488, 528
Sorodoc, Ionut-Teodor, 282, 302
Soto, Álvaro, 617
Soto, Xabier, 265, 326
Sotudeh Gharebagh, Sajad, 137, 178
Spanakis, Gerasimos, 620
Spaulding, Elizabeth, 640
Specia, Lucia, 75, 116, 330, 348, 435, 531, 636
Sperber, Matthias, 492, 526, 532, 563
Sravani, Dama, 643
Srikumar, Vivek, 155, 194, 346, 405, 552, 606
Srinet, Kavya, 341, 384
Srinivasan, Anirudh, 248, 267
Srinivasan, Balaji Vasan, 283, 302
Srinivasan, Padmini, 152, 213
Srinivasan, Soundar, 635
Srivastava, Abhishek, 74, 115
Srivastava, Himani, 627
Srivastava, Saurabh, 627
Srivastava, Shashank, 523, 585
Stüker, Sebastian, 619
Stahlberg, Felix, 526, 563
Stanojević, Miloš, 275, 313, 625
Stanovsky, Gabriel, 448, 469, 551, 561, 588, 606
Stasaski, Katherine, 349, 408, 638
Stavropoulos, Petros, 623
Steedman, Mark, 275, 313, 625
Stefanov, Peter, 46, 64
Stein, Benno, 74, 115, 232, 270, 373, 392, 471, 513
Steinert-Threlkeld, Shane, 343, 403
Steingrímsson, Steinþór, 177, 607
Steinmetz, Ina, 616
Stemle, Egon, 627, 628
Stemmer, Fabian, 641
Stenetorp, Pontus, 450, 529
Stengel-Eskin, Elias, 552, 606
Stepanov, Daniela, 469, 606
Stern, Mitchell, 153, 214
Steurer, Vanessa, 289, 329
Stewart, Robert, 622
Stockinger, Kurt, 54, 72
Stokowiec, Wojciech, 264, 326
Stone, Matthew, 11, 19, 446, 484
Stoyanov, Veselin, 427, 540, 552, 580, 588, 589
Ströbel, Marcus, 638
Strötgen, Jannik, 467, 506, 632, 633
Stratos, Karl, 346, 405
Strelec, Marek, 307, 350
Strobelt, Hendrik, 336, 396
Strope, Brian, 140, 199
Strubell, Emma, 632
Strzalkowski, Tomek, 627
Su, Chuandong, 627
Su, Enmin, 620
Su, Hui, 479, 482, 499, 519, 630
Su, Jianlin, 90, 130
Su, Jinsong, 228, 264, 482, 499, 542, 597, 644
Su, Keh-Yih, 56, 97
Su, Qi, 627
Su, Qinliang, 51, 111, 450, 528
Su, Shang-Yu, 49, 65
Su, Yu, 541, 597, 635
Subbalakshmi, Koduvayur, 622

Subbian, Karthik, 463, 584
Subramanian, Sanjay, 367, 386
Sudoh, Katsuhito, 58, 248, 273, 330, 620
Suendermann-Oeft, David, 641
Suglia, Alessandro, 523, 547
Suhane, Ayush, 64, 107
Suhara, Yoshihiko, 374, 393
Suhr, Alane, 551, 588
Sultan, Md Arafat, 128, 165, 369, 407
Sulubacak, Umut, 619
Sumita, Eiichiro, 37, 40, 99, 132, 247, 286, 335, 514
Summerville, Adam, 621
Sun, Aixin, 446, 484
Sun, Changlong, 230, 251, 287, 331
Sun, Haifeng, 270, 311
Sun, Haipeng, 247, 286
Sun, Huan, 305, 407, 544, 598, 635
Sun, Jian, 48, 65, 69, 111
Sun, Kai, 349, 368, 388, 408
Sun, Maosong, 355, 394, 427, 443, 470, 490, 501, 511
Sun, Meng, 642
Sun, Ming-Ting, 636
Sun, Mingming, 544, 582
Sun, Shuo, 238, 315, 435, 531
Sun, Simeng, 525, 587
Sun, Siqi, 377, 576
Sun, Tony, 225, 281
Sun, Weiwei, 275, 313, 452, 469, 509, 625
Sun, Weiyi, 641
Sun, Xiaobing, 244, 302
Sun, Xiaofei, 40, 99
SUN, Xu, 233, 271, 429, 490
Sun, Yu, 49, 125
Sun, Yuhui, 620
Sun, Zhiqing, 150, 190, 366, 385
Sundaram, Shiva, 157, 196
Sung, Mujeen, 250, 269
Sung, Tzu-Wei, 425, 466
Sung, Yun-hsuan, 140, 199
Sunkara, Monica, 641
Suominen, Hanna, 235, 273, 638
Surdeanu, Mihai, 159, 306, 389, 554
Susanto, Raymond Hendy, 247, 286
Suter, Benjamin, 639
Suvarna, Ashima, 375, 607
Suzuki, Jun, 40, 47, 86, 99, 227, 254, 273, 287, 325, 328, 335, 444, 483
Swanson, Ben, 376, 414
Swanson, Kyle, 367, 386
Swayamdipta, Swabha, 448, 551, 561, 606
Syed, Shahbaz, 291, 311, 471, 513
Szolovits, Peter, 353, 412, 622

Tabassum, Jeniya, 348, 390
Tachbelie, Martha Yifiru, 616
Tadepalli, Prasad, 444, 502
Tafjord, Oyvind, 535, 576
Tagliabue, Jacopo, 642
Tahiri, Mohamed-Ayoub, 643
Takahashi, Kosuke, 248, 330
Takahashi, Yujin, 58
Takamura, Hiroya, 243, 282
Takanobu, Ryuichi, 48, 125, 255, 295
Takase, Sho, 79, 137
Takayama, Junya, 36, 70
Talmina, Natalia, 366, 385
Talukdar, Partha, 305, 366, 368, 385, 500, 542, 570, 590
Tamaazousti, Youssef, 643
Tamari, Ronen, 435, 531
Tamborrino, Alexandre, 263, 324
Tan, Hongye, 54, 72
Tan, Jie, 426, 467
Tan, Liling, 247, 286
Tan, Min, 92, 132
Tan, Samson, 224, 319
Tan, Wang-Chiew, 374, 393
Tan, Xu, 33, 128, 253, 333
Tan, Yi Chern, 540, 597
Tanaka, Yu, 254, 294
Tanaka-Ishii, Kumiko, 242, 260
Tang, Chengguang, 48, 65
Tang, Duyu, 427, 429, 430, 509, 511, 512
Tang, Hao, 290, 310, 447, 464
Tang, Jie, 231, 288
Tang, Raphael, 152, 171, 211, 213, 633, 634
Tang, Shirlyn, 225, 281
Tang, Siliang, 426, 467
Tang, Yun, 170, 210
Tang, Zheng, 159, 554
Tang, Zineng, 345, 404
Tannert, Simon, 89, 110
Tao, Cui, 641
Tao, Dacheng, 93, 114
Tao, Shu, 368, 388
Tar, Chris, 140, 199
Tarasov, Alexey, 635
Tarn, Natalie, 628
Taslimipoor, Shiva, 176, 215
Tass, E. Shannon, 105, 146
Tata, Sandeep, 444, 599
Taub Tabib, Hillel, 622
Taub-Tabib, Hillel, 59, 139
Tay, Yi, 35, 155, 186, 194, 362, 382
Taylor, Andrew, 623
Taylor, Stacey, 642
Tayyar Madabushi, Harish, 628
Teich, Elke, 620
Tekiroğlu, Serra Sinem, 74, 115
Tela, Abrhalei Frezghi, 136, 491
Temčinas, Tadas, 621
Tenenbaum, Josh, 536
Tenney, Ian, 179, 216
Teshome, Getenesh, 616
Testoni, Alberto, 616, 631
Tetreault, Joel, 5, 629
Teufel, Simone, 628
Thain, Nithum, 289, 329
Thaker, Khushboo, 541, 581
Thapliyal, Ashish V., 33, 185
Thomas, Derek, 268, 308
Thomason, Jesse, 631
Thombre, Pranav, 639
Thompson, Brian, 307, 350
Thorne, James, 624
Tian, Hao, 271, 332
Tian, Junfeng, 480, 520
Tian, Xiaoyi, 119, 554
Tian, Ye, 525, 587
Tian, Yuan, 90, 130
Tian, Yuanhe, 549, 604
Tian, Zhiliang, 48, 65
Tichy, Walter F., 289, 329
Tiedemann, Jörg, 256, 337, 619
Tiedemann, Jörg, 264, 326
Tieleman, Olivier, 523, 547
Tiktinsky, Aryeh, 100, 180
Timponi Torrent, Tiago, 355, 394
Titov, Ivan, 92, 113, 354, 413, 625
Tits, Noé, 645
Tiwari, Aditya, 621
Toledo, Assaf, 471, 513
Tomashenko, Natalia, 619

Tomazic, Federico, 75, 116
Tomkins, Andrew, 35, 186
Tong, Meihan, 423, 460
Torres Rivera, Andrés, 628
Torrisi, Giovanni, 81, 140
Torroba Hennigen, Lucas, 502, 598
Toshniwal, Shubham, 363, 400, 633, 634
Touileb, Samia, 90, 130
Toutanova, Kristina, 369, 388
Toxvaerd, Flavio, 97, 117
Toyoda, Masashi, 235, 472
Tracz, Janusz, 74, 115
Trajcevski, Goce, 52, 95
Tran, Thy Thy, 502, 583
Tripathi, Aditay, 305, 368
Trischler, Adam, 155, 194, 541, 581
Trivedi, Harsh, 305, 388
Trmal, Jan, 640
Troiano, Enrica, 643
Trott, Sean, 355, 394
Tsai, Chung-Ting, 220, 277
Tsai, Emily, 353, 412
Tsarfaty, Reut, 100, 180, 492, 531, 625, 640
Tseng, Bo-Hsiang, 108, 147
Tsujii, Jun'ichi, 622
Tsuruoka, Yoshimasa, 56, 97
Tsvetkov, Yulia, 366, 403, 562, 601, 636, 643
Tu, Cunchao, 355, 394
Tu, Kewei, 236, 237, 249, 274, 276, 391, 452, 510, 626
Tu, Lifu, 173, 192
Tu, Zhaopeng, 226, 228, 263, 264
Tulkens, Stéphan, 232, 291
Turchi, Marco, 465, 562, 619, 620
Tutek, Martin, 632, 633
Twiton, Michael, 485, 505

Ubale, Rutuja, 627
Uceda-Sosa, Rosario, 75, 117
Uddin, Salah, 51, 111
Ueffing, Nicola, 642
Ul Haq, Sami, 637
Umada, Testumichi, 640
Utama, Prasetya Ajie, 569, 590
Utiyama, Masao, 37, 40, 99, 132, 247, 286
Utsuro, Takehito, 620, 624

Väth, Dirk, 418, 494
Völkel, Moritz, 418, 494
Völske, Michael, 471, 513
Vázquez, Raúl, 619
Vadakkeeveetil Sreelatha, Silpa, 507, 567
Vadapalli, Raghuram, 500, 542
Vafa, Keyon, 362, 382
Vaidya, Satyarth, 636
Vajpayee, Avijit, 627, 638
Valvoda, Josef, 321, 342, 492, 531
Van de Cruys, Tim, 165, 185
Van Durme, Benjamin, 544, 552, 559, 571, 582, 591, 598, 606, 629
Vanderlyn, Lindsey, 418, 494
Vandyke, David, 108, 147
Vania, Clara, 349, 355, 390, 394
Vanzo, Andrea, 523, 547
Varada, Venkat, 621
Varanasi, Stalin, 621
Varia, Siddharth, 565, 602
Varma, Vasudeva, 629
Varshney, Vaibhav, 627
Vashishth, Shikhar, 366, 385
Vasilescu, Bogdan, 427, 589
Vechkaeva, Anna, 623
Vechtomova, Olga, 11, 18, 352, 411, 541, 581
Venturi, Giulia, 638
Vernetti, Pedro, 616
Verspoor, Karin, 622, 623
Veselova, Eugeniia, 78, 554
Vesik, Kaili, 639
Viegas, Felipe, 546, 584
Vijayan, Priyesh, 621
Villavicencio, Aline, 616
Vincent, Julian, 90, 130
Virpioja, Sami, 256, 337
Vlachos, Andreas, 624
Vobilisetty, Sanath, 627
Voita, Elena, 132, 173
Voitot, Pascal, 263, 324
Volpi, Riccardo, 632, 633
Vorontsov, Konstantin, 78, 554
Voss, Clare, 140, 180
Vozila, Paul, 641
Vu, Ngoc Thang, 89, 110, 418, 494, 639
Vulić, Ivan, 31, 109, 175, 215, 465, 508, 562, 568, 621, 632, 633
Vyas, Aadit, 540, 597
Vylomova, Ekaterina, 639
Vázquez, Raúl, 264, 326

Wachsmuth, Henning, 232, 270, 291, 311
Wagner, Joachim, 573, 608, 626
Waibel, Alexander, 619, 620
Waites, William, 307, 350
Walker, Erin, 119, 554
Walker, Marilyn, 165, 185
Wallace, Byron, 641
Wallace, Byron C., 101, 180, 301, 366, 385, 403, 622
Wallace, Eric, 171, 211
Wallach, Hanna, 364, 401
Waltenburg, Eric, 362, 382
Waltinger, Ulli, 469, 510
WAN, Mingyu, 627
Wan, Xiaojun, 34, 67, 292, 333, 432, 433, 452, 469, 473, 474, 482, 499
Wan, Yu, 465, 526
Wandl, Florian, 640
Wang, Alex, 179, 216, 352, 411
Wang, Ante, 542, 597
Wang, Bailin, 509, 552
Wang, Baoxun, 30, 87
Wang, Bin, 620, 644
Wang, Changhan, 619
Wang, Chao, 642
Wang, Chaojun, 247, 304
Wang, Chenglong, 637
Wang, Chengyi, 252, 292
Wang, Chengyu, 250, 268
Wang, Chenyu, 231, 288
Wang, Chong, 114, 173
Wang, Danqing, 432, 473
Wang, Dong, 86, 108
Wang, Elaine Lin, 638
Wang, Fan, 31, 109
Wang, Fei, 467, 506
Wang, Feng, 53, 96, 227, 263
Wang, Fu Lee, 430, 570
Wang, Guangtao, 170, 210
Wang, Guoyin, 165, 186
Wang, Hai, 630
Wang, Haifeng, 31, 65, 109, 124, 126, 205, 271, 332, 432, 473
Wang, Hainan, 159, 375
Wang, Hanrui, 525, 601

Wang, Hao, 373, 392
Wang, Heyuan, 52, 95
Wang, Huimin, 440, 480
Wang, Jasmine, 164, 184
Wang, Jiahai, 427, 430, 509, 512
Wang, Jian, 32, 127
Wang, Jing, 559, 598
Wang, Jingjing, 251, 331
Wang, Jingyu, 270, 311
WANG, Jue, 424, 461
Wang, Kai, 233, 311, 450, 480, 520, 528
Wang, Lei, 266, 305
Wang, Liang, 36, 69
Wang, Lidan, 177, 530
Wang, Lijie, 619
Wang, Liwei, 167, 207
Wang, Longshaokan, 621
Wang, Longyue, 93, 114, 226, 263
Wang, Lu, 242, 260, 353, 363, 400, 412
Wang, Lucy, 535, 576
Wang, Lucy Lu, 350, 409
Wang, Minghan, 620
Wang, Mingxuan, 26, 179
Wang, Nan, 641
Wang, Ning, 622
Wang, Pinghui, 230, 287
Wang, Prince Zizhuang, 245, 324
Wang, Qian, 619
Wang, Quan, 429, 511
Wang, Rongbo, 627
Wang, Rui, 36, 37, 70, 132, 233, 247, 286, 291, 311, 480, 520, 563, 601
Wang, Ruili, 460, 501
Wang, Shijin, 42, 86, 108, 120
Wang, Shuai, 423, 460
Wang, Shuo, 228, 264
Wang, Sida I., 169, 209
Wang, Simi, 621
Wang, Sinong, 151, 191, 624
Wang, Taifeng, 30, 53, 86, 96
Wang, Tao, 237, 276
Wang, Tian, 642
Wang, Tianliang, 52, 95
Wang, Tianlu, 364, 401
Wang, Tianming, 433, 452, 469, 474, 482, 499
Wang, Tong, 541, 581
Wang, Wei, 52, 213, 525, 587
Wang, Wenhui, 450, 488
Wang, Wenlin, 165, 186
Wang, William Yang, 33, 185, 225, 245, 281, 324, 541, 597, 631
Wang, Xiaobin, 449, 486
WANG, Xiaojie, 443, 483
Wang, Xiaolan, 374, 393
Wang, Xiaoyan, 230, 287
Wang, Xiaoyang, 68, 129
Wang, Xin, 631
Wang, Xing, 226, 263
Wang, Xinyi, 562, 601
Wang, Xinyu, 237, 276, 626
Wang, Xuan, 101, 295
Wang, Yan, 266, 305, 422, 479
Wang, Yanjie, 155, 194
Wang, Yining, 619
Wang, Yiyi, 638
Wang, Yonggang, 549, 604
Wang, Yu, 179, 216, 445, 502
Wang, Yuanjie, 644
Wang, Yuanzhuo, 430, 511
Wang, Yue, 90, 130
Wang, Yuping, 26, 179
Wang, Yuquan, 231, 288
Wang, Yuxia, 622
Wang, Yuxuan, 26, 179
Wang, Zhen, 544, 598
Wang, Zheng, 444, 461
Wang, Zhenyi, 68, 129
Wang, Zhiguo, 306, 369
Wang, Zhiruo, 427, 509
Wang, Ziang, 46, 64
Wang, Zihao, 228, 326
Wang, Ziyang, 246, 285
Wang, Zongsheng, 30, 87
Ward, Todd, 75, 117
Warren, Christopher, 226, 346
Warstadt, Alex, 569, 590
Waszczuk, Jakub, 628
Watanabe, Shinji, 436, 495
Wattenhofer, Roger, 177, 572
Way, Andy, 177, 265, 326, 607, 636
Webster, Kellie, 364, 402
Wedekind, Jürgen, 639
Wei, Daimeng, 620
Wei, Furu, 450, 488
Wei, Jielong, 484, 523
Wei, Penghui, 232, 270
Wei, Xiangpeng, 38, 133
Wei, Yang, 236, 313
Wei, Yizhen, 620
Wei, Zhenkai, 251, 311
Wei, Zhepei, 90, 130
Wei, Zhongyu, 424, 462
Weigelt, Sebastian, 289, 329
Wein, Shira, 617
Weiss, Gail, 40, 198
Weiss, Jeremy, 540, 597
Welbl, Johannes, 632
Weld, Daniel, 152, 213, 350, 409
Weld, Daniel S., 346, 405
Welleck, Sean, 341, 384
Weller, Orion, 105, 146, 628
Weller-Di Marco, Marion, 286, 304
Wen, Haoyang, 450, 528
Wen, Tsung-Hsien, 621
Wen, Zheng, 165, 186
Wen, Zhongzhen, 36, 69
Wendt, James Bradley, 444, 599
Weng, Rongxiang, 38, 133
Weng, Wei-Hung, 622
Wenzek, Guillaume, 552, 589
West, Rebecca, 642
West, Robert, 93, 133
Weston, Jason, 147, 164, 205, 206, 341, 348, 384, 408
White, Aaron Steven, 155, 194, 552, 606
White, Jennifer, 639
White, Max, 638
White, Ryen, 566, 603
Whitehead, Spencer, 140, 167, 180, 207
Wiechmann, Daniel, 638
Wiegreffe, Sarah, 301, 403
Wiemerslage, Adam, 638
Wiesner, Matthew, 307, 371, 640
Wieting, John, 501, 544
Wiggins, Dion, 307, 350
Wijaya, Derry Tanti, 566, 603
Wilcox, Ethan, 105, 120, 199, 203
Wiley, Korah, 638
Wilken, Patrick, 619, 620
Williams, Adina, 321, 342, 348, 408, 449, 486, 492, 531, 569, 590
Williams, Philip, 92, 113, 620
Williamson, Mary, 147, 205

Willis, Angelica, 633, 634
Wilmot, David, 105, 204
Winata, Genta Indra, 30, 108, 252, 333, 633, 634
Wiseman, Sam, 173, 192, 346, 405
Woldemariyam, Kidane, 616
Wolf, Adva, 464, 505
Wolfson, Tomer, 305, 367, 368, 386
Wong, Derek F., 30, 38, 87, 93, 465, 526
Wong, Kam-Fai, 242, 260, 440, 480
Wong, KayYen, 425, 465
Wonsever, Dina, 625
Wu, Bo, 140, 180
Wu, Bowen, 30, 87
Wu, Chen, 245, 324
Wu, Chien-Sheng, 55, 72
Wu, Chuhan, 226, 248, 307, 324
Wu, Dapeng, 169, 209
Wu, Fangzhao, 52, 95, 226, 248, 307, 324
Wu, Fei, 40, 99, 423, 467, 501, 506
Wu, Hao, 228, 326
Wu, Hongtao, 248, 307
Wu, Hua, 31, 65, 109, 124, 126, 205, 271, 332, 432, 473, 482, 499, 644
Wu, Jiele, 252, 292
Wu, Jiemin, 430, 570
Wu, John, 322, 343
Wu, Joy, 622
Wu, Lianwei, 64, 107
Wu, Lingfei, 353, 368, 388, 412, 543, 565, 582, 602
Wu, Liting, 644
Wu, Qi, 631
Wu, Qianhui, 445, 502
Wu, San He, 642
Wu, Shijie, 427, 527, 550, 588, 633, 634, 639
Wu, Shu, 36, 69
Wu, Sixing, 422, 479
Wu, Tongshuang, 348, 390
Wu, Wei, 467, 506
Wu, Wenhao, 432, 473
Wu, Winnie, 248, 307
Wu, Winston, 637
Wu, Xiao-Ming, 66, 126
Wu, Xiaoting, 480, 520
Wu, Xiuyu, 636
Wu, Yonghui, 173, 192
Wu, Youzheng, 79, 138
Wu, Yu, 87, 108, 246, 252, 285, 292
Wu, Yuanbin, 236, 313
Wu, Yunfang, 429, 490, 636
Wu, Yuting, 444, 461
Wu, Zhanghao, 525, 601
Wu, Zhiyong, 282, 322
Wu, Zhonghai, 422, 479
Wuebker, Joern, 92, 113

Xia, Chen, 449, 605
Xia, Fei, 549, 604, 641
Xia, Mengzhou, 566, 603, 643
Xia, Patrick, 544, 582
Xia, Rui, 232, 242, 299, 331
Xiang, Beilei, 640
Xiang, Bing, 306, 369
Xiang, Rong, 627
Xiang, Yang, 643
Xiao, Chaojun, 355, 394
Xiao, Lin, 230, 288
Xiao, Tong, 231, 246, 285, 288, 447, 485, 637
Xiao, Xinyan, 271, 332, 432, 473, 482, 499
Xiao, Zhou, 479, 519
Xie, Haoran, 430, 570
Xie, Hongtao, 429, 511
Xie, Jing Yi, 640
Xie, Jun, 246, 285, 423, 460
Xie, Ning, 620
Xie, Pengjun, 69, 111, 449, 486
Xie, Ruobing, 443, 483
Xie, Xing, 52, 95, 248, 287, 307, 328
Xie, Yuxi, 89, 129
Xie, Zhiwen, 424, 461
Xin, Ji, 152, 171, 211, 213
Xing, Xinyu, 432, 473
Xing, Yuqing, 442, 459
Xiong, Caiming, 55, 72, 301, 337, 364, 377, 385, 401, 540, 597, 621
Xiong, Chenyan, 147, 184, 490, 511
Xiong, Deyi, 37, 113, 114, 247, 304, 480, 520
Xiong, JinJun, 565, 602
Xiong, Yuxuan, 36, 70
Xu, Benfeng, 429, 511
Xu, Bin, 423, 460
Xu, Boyan, 35, 129
Xu, Canwen, 34, 110, 248, 307
Xu, Dongfang, 559, 582
Xu, Frank F., 427, 589
Xu, Guangwei, 69, 111, 449, 486
Xu, Haiyang, 36, 70
Xu, Hongfei, 37, 113, 114, 247, 304
Xu, Hongzhi, 449, 486
Xu, Hua, 252, 292
Xu, Jiacheng, 352, 411
Xu, Jingjing, 430, 512
XU, Jitao, 92, 113
Xu, Jun, 124, 205
Xu, Ke, 450, 488
Xu, Kun, 363, 400, 542, 597
Xu, Linli, 37, 113
Xu, Liyan, 622
Xu, Minjie, 515, 593
Xu, Nan, 253, 292
Xu, Nuo, 230, 287, 447, 485
Xu, Peng, 30, 108, 252, 333
Xu, Ruifeng, 251, 291
Xu, Runxin, 26, 179
Xu, Ruochen, 566, 603
Xu, Siyong, 287, 328
Xu, Song, 79, 138
Xu, Wei, 149, 189, 348, 390, 541, 581
Xu, Weidi, 53, 96
Xu, Weijia, 617
XU, Weiran, 48, 125
Xu, Wentao, 263, 324
Xu, Xiao, 440, 519
Xu, Xiaofei, 232, 270
Xu, Xiaoxiao, 227, 263
Xu, Xinnuo, 353, 412
Xu, Zenan, 430, 512
Xue, Haiyang, 644
Xue, Kang, 638
Xue, Lanqing, 48, 65
Xue, Zhengshan, 619

Yadav, Vikas, 306, 389
Yahav, Eran, 40, 198
Yalta, Nelson, 436, 495
Yan, Chenwei, 436, 515
Yan, Guangfeng, 66, 126
Yan, Hang, 236, 274, 460, 522
Yan, Ming, 488, 529
Yan, Nian, 642
Yan, Rui, 422, 519
Yan, Shengjia, 467, 506
Yan, Yu, 451, 488

Yan, Yuanmeng, 48, 125
Yanai, Kohsuke, 234, 272
Yanaka, Hitomi, 429, 490
Yaneva, Victoria, 638
Yang, Baosong, 465, 526
Yang, Changbing, 640
Yang, Changlin, 426, 467
Yang, Charles, 449, 486
Yang, Cheng, 287, 328
Yang, Chenghao, 427, 470, 630, 633, 634
Yang, Chenyu, 452, 470
Yang, Ching-Yu, 220, 277
Yang, Diyi, 150, 190
Yang, Grace Hui, 349, 408
Yang, Hao, 620
Yang, Jian, 425, 465
Yang, Kaicheng, 252, 292
Yang, Li, 242, 299
Yang, Liang, 627
Yang, Michael, 636
Yang, Min, 232, 251, 270, 291
Yang, Muyun, 644
Yang, Sohee, 47, 147
Yang, Yan, 440, 480
Yang, Yating, 98, 453
Yang, Yi, 52, 95
Yang, Yiming, 128, 150, 186, 190, 366, 385, 566, 603
Yang, Yinfei, 140, 199, 525, 587
Yang, Yujiu, 39, 431
Yang, Yunyi, 233, 311
Yang, Zhenglu, 252, 292
Yang, Zhengyuan, 228, 264
Yang, Zichao, 150, 190
Yang, Ziqing, 42, 120
Yannakoudakis, Helen, 152, 213, 288, 328, 638
Yao, Bingsheng, 630
YAO, Jianmin, 251, 311
Yao, Kaisheng, 31, 125, 443, 483
Yao, Lili, 638
Yao, Shaowei, 292, 333, 482, 499
Yao, Ting, 266, 305
Yao, Yuekun, 620
Yarlott, W. Victor, 629
Yarowsky, David, 640
Yatskar, Mark, 356, 395
Yazdi, Ram, 170, 210
Ye, Chenchen, 36, 70
Ye, Mao, 245, 405
Ye, Qinyuan, 555, 593
Ye, Xi, 429, 569
Ye, Yajie, 275, 313
Ye, Yunming, 232, 270
Ye, Zhiquan, 227, 263
Yeo, Catherine, 617
Yeres, Phil, 179, 216
Yi, Yanzhi, 56, 97
Yih, Scott Wen-tau, 635
Yih, Wen-tau, 12, 22, 552, 589, 624
Yin, Biao, 321, 342
Yin, Da, 251, 311, 356, 395
Yin, Dawei, 440, 479
Yin, Fan, 243, 302
Yin, Jian, 263, 324, 427, 429, 430, 450, 509, 511, 512, 528
Yin, Pengcheng, 427, 552, 589
Yin, Xiang, 26, 179
Yin, Yongjing, 228, 264
Yoder, Charlotte, 639
Yogatama, Dani, 321, 367, 492, 531
Yokoi, Sho, 235, 444, 483, 491
Yoon, Jooyoung, 98, 453
Yoon, Seunghyun, 52, 95
Yoshida, Issei, 51, 69
Yoshinaga, Naoki, 235, 472
Yoshino, Koichiro, 58, 273
You, Daniel, 639
You, Weiqiu, 525, 587
Youssef, Abdou, 628
Yu, Bingqing, 642
Yu, Bowen, 245, 284
Yu, Changlong, 250, 309
Yu, Cong, 636
Yu, Dian, 349, 368, 388, 408, 450, 529
Yu, Dong, 68, 129, 167, 207, 349, 363, 368, 388, 400, 408, 450, 529, 542, 597
Yu, Heng, 38, 93, 133
Yu, Hong, 160, 357
Yu, Hongkun, 150, 190
Yu, Jianfei, 232, 242, 299, 331
Yu, Jianxing, 450, 480, 520, 528
Yu, Jifan, 231, 288
Yu, Juntao, 444, 483
Yu, Kai, 50, 67, 430, 452, 470, 511
Yu, Lei, 264, 326
Yu, Licheng, 547, 586
Yu, Lili, 169, 209, 367, 386
Yu, Miaomiao, 39, 431
Yu, Mo, 630
Yu, Qian, 35, 68
Yu, Seunghak, 418, 515
YU, SZ-HAN, 175, 215
Yu, Wenhao, 368, 388
Yu, Wenmeng, 252, 292
Yu, Xiang, 89, 110, 639
Yu, Xueli, 36, 69
Yu, Yaoliang, 152, 171, 211, 213
Yu, Youngjae, 627
Yu, Zhou, 48, 86, 108, 109, 440, 480, 621
Yuan, Arianna, 467, 506
Yuan, Chaofa, 251, 291
Yuan, Fei, 54, 72
Yuan, Peng, 79, 138
Yuan, Quan, 231, 289
Yuan, Steve, 140, 199
Yuan, Xingdi, 155, 194, 541, 581
Yue, Xiang, 305, 407
Yuma, Tsuta, 235, 472
Yun, Shuang, 284, 325
Yuret, Deniz, 632, 633

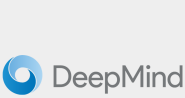
Zad, Samira, 629
Zadeh, AmirAli Bagher, 645
Zaidi, Mohd Abbas, 619
Zalmout, Nasser, 549, 604
Zang, Xiaoxue, 621
Zang, Yuan, 427, 470
Zaragoza, Jaume, 307, 350
Zareian, Alireza, 140, 167, 180, 207
Zarrieß, Sina, 446, 585
Zayed, Omnia, 627
Zeman, Daniel, 625
Zeng, Jichuan, 337, 377
Zeng, Jiehang, 447, 485
Zeng, Jingjie, 627
Zeng, Qi, 167, 207
Zeng, Qingkai, 368, 388
Zeng, Xingshan, 242, 260
Zeng, Ying, 26, 179
Zeng, Zhixiong, 253, 292
Zenkel, Thomas, 92, 113
Zesch, Torsten, 638
Zettlemoyer, Luke, 57, 166, 175, 187, 346, 405, 427, 469, 540, 551, 552, 580, 588, 589, 606

Zewoudie, Abraham Woubie, 136, 491
Zha, Yefei, 48, 124
Zhang, Biao, 92, 113
Zhang, Bo, 233, 291
Zhang, Boliang, 563, 601
Zhang, Bowen, 232, 270
Zhang, Byoung-Tak, 631
Zhang, Chen, 253, 333
Zhang, Cheng, 440, 479
Zhang, Chuanqiang, 644
Zhang, Dakun, 637
Zhang, Dawei, 422, 479
Zhang, Dongdong, 246, 285, 425, 465
Zhang, Dongmei, 87, 109
Zhang, Guanhua, 281, 319
Zhang, Haisong, 50, 110, 250, 309
Zhang, Hanchu, 642
Zhang, Hao, 446, 484, 488, 529
Zhang, Haoran, 565, 602
Zhang, Haoyu, 69, 111
Zhang, Hongming, 371, 409
Zhang, Houyu, 147, 184
Zhang, Ji, 234, 272
Zhang, Jiajun, 79, 137, 619, 644
Zhang, Jian, 169, 209
Zhang, Jiarui, 644
Zhang, Jinchao, 440, 519
Zhang, Jingyi, 37, 113, 114
Zhang, Jingyuan, 544, 582
Zhang, Jinhui, 627
Zhang, Jipeng, 266, 305
Zhang, Jun, 227, 263
Zhang, Junqi, 281, 319
Zhang, Justine, 361, 381
Zhang, Kunpeng, 52, 95
Zhang, Licheng, 429, 511
Zhang, Longyin, 442, 459
Zhang, Meishan, 459, 469, 509, 521
Zhang, Meng, 427, 470
Zhang, Min, 92, 132, 233, 236, 251, 275, 291, 331
Zhang, Ming, 422, 519
Zhang, Mozhi, 151, 191
Zhang, Nevin L., 48, 65
Zhang, Qi, 424, 462
Zhang, Qian, 619
Zhang, Qiong, 35, 68, 230, 287
Zhang, Rong, 75, 117, 233, 272
Zhang, Rui, 49, 125, 540, 597
Zhang, Ruiqing, 644
Zhang, Ruiyi, 165, 186
Zhang, Shaohua, 53, 96
Zhang, Sheng, 552, 606
Zhang, Shengming, 52, 95
Zhang, Sijia, 644
Zhang, Tao, 444, 461
Zhang, Tianyang, 355, 394
Zhang, Tong, 549, 604
Zhang, Wei-Nan, 422, 479
Zhang, Wenzheng, 230, 328, 445, 502
Zhang, Xiao, 625
Zhang, Xiaotong, 66, 126
Zhang, Xiaoyi, 355, 394
Zhang, Xijin, 26, 179
Zhang, Yi, 233, 271, 299, 382, 621
Zhang, Yichi, 48, 109
Zhang, Yifan, 418, 515
Zhang, Yizhe, 165, 186, 377, 504, 560, 576
Zhang, Yong, 228, 326
Zhang, Yongdong, 429, 511
Zhang, Yu, 236, 275
Zhang, Yuan, 547, 585, 619
Zhang, Yuanzhe, 444, 461
Zhang, Yue, 13, 38, 87, 92, 108, 132, 133, 233, 290, 291, 310, 363, 400, 423, 440, 447, 459, 461, 504, 519, 521, 542, 597
Zhang, Yueping, 54, 72
Zhang, Yufeng, 36, 69
Zhang, Yuhao, 160, 216, 353, 412, 447, 485
Zhang, Yuhui, 160, 216, 351, 375, 633, 634
Zhang, Yunyi, 546, 584
Zhang, Zeyu, 559, 582
Zhang, Zheng, 47, 124, 255, 295
Zhang, Zhisong, 501, 582
Zhang, Zusheng, 430, 570
Zhao, Chao, 165, 185
Zhao, Dongyan, 444, 461
Zhao, Hai, 443, 522
Zhao, He, 233, 272
Zhao, Hongyan, 54, 72
Zhao, Jason, 624
Zhao, Jiahao, 232, 270
Zhao, Jieyu, 52, 213, 224, 225, 281, 401
Zhao, Jun, 230, 288, 436, 444, 461, 515
Zhao, Junzhou, 230, 287
Zhao, Qi, 444, 599
Zhao, Sanqiang, 479, 519
Zhao, Tiancheng, 60, 147, 160, 205
Zhao, Tianyu, 30, 86
Zhao, Tiejun, 247, 281, 286, 319, 644
Zhao, Tong, 624
Zhao, Tuo, 114, 150, 173, 190
Zhao, Wei, 79, 93, 133, 138
Zhao, Xiaofang, 440, 479
Zhao, Xinran, 371, 409
Zhao, Xu, 228, 326
Zhao, Yanbin, 50, 67, 430, 452, 470, 511
Zhao, Yang, 619
Zhao, Yinggong, 30, 124
Zhao, Yiyun, 342, 386
Zhao, Yu, 443, 483
Zhao, Yuanyuan, 452, 469
Zhao, Zhe, 427, 509
Zhao, Zhihao, 540, 580
Zhao, Zhou, 33, 128, 253, 333
Zheleva, Elena, 630
Zhelezniak, Vitalii, 551, 588
Zheng, Baigong, 38, 174, 192, 193
Zheng, Bo, 450, 528
Zheng, Changmeng, 484, 523
Zheng, Che, 35, 186
Zheng, Chen, 523, 585
Zheng, Chujie, 479, 519
Zheng, Emily, 365, 402
Zheng, Guangtao, 367, 403
Zheng, Haitao, 449, 486
Zheng, Lin, 51, 111
Zheng, Renjie, 38, 174, 192, 193
Zheng, Shun, 263, 324
Zheng, Stephan, 540, 597, 621
Zheng, SuHang, 227, 263
Zheng, Xiaoqing, 447, 485
Zheng, Xinyi, 635
Zheng, Yining, 432, 473
Zheng, Zaixiang, 58, 375
Zhong, Haoxi, 355, 394
Zhong, Lei, 46, 64
Zhong, Ming, 432, 473
Zhong, Qingyang, 231, 288
Zhong, Ruiqi, 153, 214
Zhong, Ting, 52, 95
Zhong, Wanjun, 427, 430, 509, 512
Zhong, Xu, 623

Zhong, Yang, 541, 581
Zhong, Yu, 364, 402
Zhou, Ben, 512, 570
Zhou, Bin, 87, 109
Zhou, Bowen, 79, 138, 170, 210
Zhou, Chulun, 228, 264, 482, 499, 644
Zhou, Denny, 150, 190
Zhou, Deyu, 36, 70, 260, 299
Zhou, Fan, 52, 95
Zhou, Guangyou, 424, 461
Zhou, Guodong, 251, 331, 442, 459
Zhou, Hao, 26, 35, 89, 179, 245, 284, 479, 519
Zhou, Jiawei, 132, 192
Zhou, Jie, 35, 69, 89, 111, 228, 264, 286, 440, 443, 479, 501, 519
Zhou, Jin, 566, 603
Zhou, Joey Tianyi, 353, 412, 446, 484, 488, 529
Zhou, Li, 148, 206
Zhou, Linqi, 249, 391
Zhou, Long, 619, 644
Zhou, Mantong, 266, 305
Zhou, Ming, 87, 108, 246, 248, 252, 285, 287, 292, 307, 328, 425, 427, 429, 430, 450, 465, 509, 511, 512, 528
Zhou, Peng, 427, 509
Zhou, Qiji, 290, 310, 447, 464
Zhou, Shuyan, 501, 544, 545, 599
Zhou, Wenjie, 429, 490
Zhou, Xiang, 570, 591
Zhou, Xin, 547, 585
Zhou, Xu, 499, 580
Zhou, Xuhui, 58, 375
Zhou, Yang, 422, 479, 624
Zhou, Yi, 447, 460, 485, 501
Zhou, Yichao, 52, 213
Zhou, Yikai, 465, 526
Zhou, Yu, 79, 137
Zhou, Zhengping, 633, 634
Zhou, Zhihan, 86, 108
Zhu, Changfeng, 93, 133
Zhu, Conghui, 281, 319
Zhu, Jingbo, 246, 285, 447, 485, 637
Zhu, Junnan, 79, 137
Zhu, Ligeng, 525, 601
Zhu, Lixing, 260, 299
Zhu, Muhua, 449, 486
Zhu, Qi, 47, 124, 255, 295, 349, 408
Zhu, Qile, 169, 209
Zhu, Song-Chun, 446, 586
Zhu, Su, 50, 67, 430, 452, 470, 511
Zhu, Wang, 167, 207
Zhu, Xiaodan, 48, 65, 69, 111, 499, 580
Zhu, Xiaoyan, 47, 124, 255, 295, 479, 519, 540, 580
Zhu, Xueyun, 179, 216
Zhu, Yilin, 252, 292
Zhu, Yutao, 566, 603
Zhuang, Yimeng, 619
Zhuang, Yuan, 351, 607
Zhuang, Yueting, 426, 467
Zmigrod, Ran, 321, 342
Zong, Chengqing, 79, 137, 619, 644
Zong, Shi, 361, 381
Zou, Bowei, 251, 311
Zou, James, 86, 108
Zou, Jiyun, 252, 292
Zuidema, Willem, 283, 322
Zuylen, Madeleine van, 502, 599

Diamond Sponsors

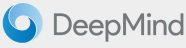
Platinum Sponsors



Gold Sponsors

Silver Sponsors

Bronze Sponsors

Supporter Sponsors

Diversity & Inclusion Champion Sponsors