NAACL 2022

10-15 July | Seattle

Contents

1

Conference Information

Message from the General Chair

Welcome to the 2022 meeting of the North American Association for Computational Linguistics! Due to the COVID-19 pandemic, NAACL-2021 was held virtually, and NAACL-2022 is the first major NLP conference that is run as an hybrid conference in North America. It is my pleasure to welcome many of you who are joining us in Seattle, as well as those who chose to participate in the conference virtually.

COVID safety is important to us and we will do whatever we can to help you enjoy the in-person conference despite the difficulties we all experience coming back to normality. At the same time, thanks to the virtual conference platform put together by Underline, we hope that our virtual attendees will experience the conference almost as if they are in Seattle and enjoy the conference.

NAACL-2022 decided, along with ACL-2022, to experiment with a new reviewing process, based on "rolling review" (ARR). While we believe that, eventually, this process will converge to an efficient review process that would benefit our community, pioneering such a process is not without difficulties. This would not have been possible without the incredible effort, devotion, thoughtfulness, patience, and many work hours put by our program chairs, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, and the help from the ACL-2022 program chairs, Smaranda Muresan, Preslav Nakov, Aline Villavicencio. This process necessitated developing a new software package to support our publication, an effort that was done in collaboration with ACL-2022, and I am thankful to Ryan Cotterell who led this effort.

Among other innovations we installed in NAACL-2022 is a reproducibility track, where we attempted to incentivize authors to release models, code, and other information necessary to reproduce the main results and findings of their papers. We hope that this effort, led by Niranjan Balasubramanian, Jesse Dodge, Annie Louis, Daniel Deutsch and Yash Kumar Lal, will be followed in future conferences. Other initiatives include incorporating a "Responsible NLP Research" checklist into the submission process, a new special theme on "Human-Centered Natural Language Processing", and many innovative activities led by our very active and thoughtful Diversity and Inclusion Committee, led by Diana Galván, Snigdha Chaturvedi and Yonatan Bisk, with Pranav A and Luciana Benotti as advisors.

Organizing a conference as large as NAACL, especially under the constraints of the times we live in, requires the support of a large number of volunteers who care deeply about our community and are willing to spend a lot of time and effort in this long process. It is an honor to coordinate such a team. I would like to thank the members of the organizing committee for their dedication, creativity, and hard work.

First, it is hard to imagine the amount of thought, care, and time, our program chairs Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz put into all aspects of organizing this conference – resulting in an exciting and high quality scientific program.

Many other volunteers have worked hard to make this conference a success and it would not be possible to name all of them here. I will only list the chairs of the main committees whose dedication, creativity, hard work and lively communication contributed to making NAACL-2022 a successful event:

- The diversity & inclusion committee chaired by Diana Galván, Snigdha Chaturvedi and Yonatan Bisk, with advisors Pranav A and Luciana Benotti.
- The industry track chairs, Rashmi Gangadharaiah, Anastassia Loukina and Bonan Min, and advisors Owen Rambow and Yunyao Li.
- The tutorial chairs, Cecilia Alm, Miguel Ballesteros and Yulia Tsvetkov.
- The demonstration chairs, Hannaneh Hajishirzi, Qiang Ning and Avi Sil.
- The workshop chairs, Dan Goldwasser, Yunyao Li and Ashish Sabharwal.
- The student research workshop chairs, Daphne Ippolito, Liunian Harold Li, Maria Leonor Pacheco and advisors, Danqi Chen and Nianwen Xue.
- The publication chairs, Ryan Cotterell, Danilo Croce and Jordan Zhang.
- The reproducibility chairs, Niranjan Balasubramanian, Jesse Dodge, Annie Louis, Daniel Deutsch and Yash Kumar Lal.
- The sponsorship chair, Byron Wallace.
- The volunteer chair, Daniel Khashabi.
- The publicity chairs, Nanyun (Violet) Peng, Emily Sheng, Sameer Singh.
- The virtual infrastructure chairs, Deepak Ramachandran, Martín Villalba, and Rishita Anubhai.
- The website chairs, Ice Pasupat and Vered Shwartz.

Many thanks to Chris Callison-Burch, the ACL Sponsorship Director, for helping the NAACL-2022 Sponsorship chair, Byron Wallace, managing the relations between the sponsors and NAACL-2022.

I am also very grateful to the chairs of previous years' conferences, who were always ready to help and share their experience, and to the members of the ACL and NAACL Executive Committees for their support, feedback and advice.

As usual, special thanks go to Priscilla Rasmussen and to Jennifer Rachford who has stepped into the role of the ACL business manager just in time to help us with NAACL-22. They have been our local organizers and have dealt with all aspects of organizing and managing the conference, from room assignment, to food, to COVID tests.

Finally, I would like to thank all authors, invited speakers and panelists, area chairs and reviewers, the volunteers organizing and chairing sessions, and all attendees, in-person and virtual, for making this a scientifically exciting and socially engaging conference.

Welcome and hope you all enjoy the conference!

Dan Roth
University of Pennsylvania and AWS AI Labs
NAACL-2022 General Chair
June 2022

Message from the Program Chairs

Welcome to the 2022 Annual Conference of the North American Association for Computational Linguistics! For the first time, NAACL-HLT 2022 is a hybrid conference. After two years of exclusively virtual conferences due to the COVID-19 pandemic, we are pleased that attendees and presenters will be able to join us in person in Seattle and from all over the world online for this year's edition.

Review Process NAACL 2022 invited the submission of long and short papers featuring substantial, original, and unpublished research in all aspects of Computational Linguistics and Natural Language Processing (NLP). Our paper review process was organized in a hierarchical structure similar to recent years. We recruited 62 senior area chairs (SACs) for 26 areas, following the areas defined for NAACL 2022. There were two paths for submitting papers: special theme papers were directly submitted to the NAACL OpenReview site, and other main conference papers were reviewed through a new ACL-wide centralized reviewing process. In coordination with the ACL 2022 organizers, we experimented with the ACL Rolling Review (ARR) introduced as part of an initiative to improve efficiency and turnaround of reviewing for ACL conferences. Within this system, reviewing and acceptance of papers to publication venues was done in a two-step process: (1) centralized rolling review via ARR, where submissions receive reviews and meta-reviews from ARR reviewers and action editors; (2) commitment to a publication venue (e.g., NAACL 2022), so that Senior Area Chairs and Program Chairs make acceptance decisions for a submission using the ARR reviews and meta-reviews. During the first phase of the review process, we served as guest Editors in Chief for the ACL Rolling Review and worked to ensure that all papers submitted received at least three review and one meta-review, while balancing the reviewing load for reviewers and action editors. NAACL SACs acted as guest senior area chairs in the ARR system, by helping monitor review progress and supporting the 408 action editors and 3379 reviewers in their work. While the new reviewing mechanism was not as smooth as one could have hoped for, all papers submitted to ARR received at least three reviews and a meta-review, so that authors could decide to commit it to NAACL 2022 if they wanted to. The ACL Executive Committee, based on feedback from the community, will decide whether the advantages of a centralized rolling review system outweigh the disadvantages, taking into account the fast growth of our research field. Once papers were committed to the NAACL OpenReview site, SACs were in charge of making acceptance recommendation per area, taking into account the submission itself, (meta-)reviews, as well as comments to SACs provided by the authors and ethics reviews when applicable. In coordination with Jesse Dodge, Anna Rogers, Margot Mieskes, Amanda Stent, and the ACL Ethics Committee, we incorporated a "Responsible NLP Research" checklist into the submission process, designed to encourage best research practices in our field, from an ethics and reproducibility perspective. The ARR Responsible NLP Research checklist is largely based on the NeurIPS 2021 paper checklist, the reproducible data checklist from Rogers, Baldwin, Leins's paper "Just What do You Think You're Doing, Dave? A Checklist for Responsible Data Use in NLP", and the NLP Reproducibility checklist introduced by Dodge, Gururangan, Card, Schwartz and Smith in "Show Your Work: Improved Reporting of Experimental Results". Authors were asked to follow the ACL code of ethics and to fill the checklist to ensure that best practices are put in place. Reviewers were asked to consult the checklist when deciding whether the paper requires ethics review. Based on input from reviewers and action editors, SACs flagged papers that required an in-depth ethics review, which was handled by a committee of 11 ethics reviewers. The ethics chairs provided guidance and office hours to help SACs decide when ethics review was required. The ethics reviews were integrated in the final acceptance recommendation by SACs and decisions by PCs.

Special Theme We highlighted "Human-Centered Natural Language Processing" as the special theme for the conference. As NLP applications increasingly mediate people's lives, it is crucial to understand how the design decisions made throughout the NLP research and development lifecycle impact people, whether there are users, developers, data providers or other stakeholders. For NAACL 2022, we invited submissions that address research questions that meaningfully incorporate stakeholders in the design, development, and evaluation of NLP resources, models and systems. We particularly encouraged submissions that bring together perspectives and methods from NLP and Human-Computer Interaction. Given their interdisciplinary nature, theme papers were reviewed through a dedicated process by reviewers with expertise in NLP and in Human-Computer Interaction. We received 52 submissions to the special theme, of which 14 have been accepted to appear at the conference.

Submission Statistics The ACL Rolling Review received 196 submissions in December and 1897 in January, which were the two submission deadlines between the ACL and NAACL commitment deadlines. Of these 2103 submissions, 56% (1073) were committed to NAACL 2022 for the senior program committee to make an acceptance decision. We accepted a total of 442 papers (358 long papers and 84 short papers), representing 21.96% of papers submitted to ARR in December and January and to the NAACL special theme, and 41.19% of papers committed to NAACL (including the special theme papers). As a reference point, NAACL-HLT 2021 received 1797 submissions and accepted 477 papers, including 350 long and 127 short, for an overall acceptance rate of 26%. From the accepted papers, and based on the nominations from SACs, the best paper committee selected best papers, as well as a small number of outstanding papers with the goal of recognizing diverse types of contributions (including contributions to the special theme on human-centered NLP; innovation in model design, training or evaluation; resource or dataset contribution).

Additionally, 209 submissions (183 long and 26 short) were accepted for publications in the "Findings of ACL: NAACL 2022" (or Findings for short), an online companion publication for papers that are not accepted for publication in the main conference, but nonetheless have been assessed by the program committee as solid work with sufficient substance. A total of 5 accepted Findings papers were withdrawn. Findings paper were given the option to be presented as posters during the main conference: 183 took this opportunity and will be presented either in person or virtually.

NAACL 2022 will also feature 15 papers that were published at Transactions of the Association for Computational Linguistics (TACL) and 3 papers from the journal of Computational Linguistics (CL).

Program Format The conference program was designed to allow for presentation and attendance in person in Seattle and virtually from all over the world. Oral sessions will consist of presentations done either in person or virtually. The Q&A session for each paper will alternate between in-person and online questions, with a volunteer helping monitor the online questions. All oral sessions will be live-streamed and recorded. All main conference posters will be presented with a 5-minute video pitch available online and with a virtual Q&A session, where papers will be grouped by topic to foster discussion. In addition, authors who attend the conference in Seattle will present their poster in person during traditional poster sessions. Finally, asynchronous interaction between authors and attendees will be made possible before, during and after the conference on the Underline platform. We also chose to start the conference early in the morning to overlap with normal waking hours in distant time zones.

The program includes several plenary sessions, which we hope will provide thought-provoking perspectives that will enrich discussions during the conference and beyond. In addition to a session for best paper awards, we are delighted to have keynote talks by Batya Friedman (University of Washington) and Manuel Montes-y-Gómez (National Institute of Astrophysics, Optics and Electronics of Mexico). Dan Roth (University of Pennsylvania and Amazon) will moderate a discussion on the role of linguistics and symbolic representations in NLP, with panelists Chitta Baral (Arizona State University), Emily Bender (University of Washington), Dilek Hakkani-tur (Amazon), and Christopher D. Manning (Stanford University). The industry track, demonstrations track and the student research workshop will have dedicated sessions during the main conference to round up the program, including a plenary panel on careers in NLP organized by the industry track chairs.

Gratitude NAACL would not have been possible without the hard work of many volunteers. We are very grateful to all who contributed to make the conference possible, especially given the ongoing challenges raised by the COVID-19 pandemic.

We would like to start by thanking all the authors who submitted their work to the ACL Rolling Review and NAACL 2022. We could only accept a small fraction of submissions but hope that most papers received valuable feedback.

Next, we would like to thank all members of the Program Committee – who are too numerous to be all named here, but are listed elsewhere in the proceedings:

- The senior area chairs, who were incredibly responsive throughout the reviewing process and patiently helped improve the new reviewing infrastructure.
- ARR action editors and reviewers. Special thanks to those who stepped in at the last minute to serve as emergency reviewers. This was tremendously appreciated!
- The special theme area chair, Jeff Bigham, and all reviewers, with a special note of appreciation for those who contributed their time and expertise even though they do not usually publish in NLP conferences.
- The ethics Chairs, Kai-Wei Chang, Dirk Hovy and Diyi Yang, for designing a process to encourage consistent evaluation of ethical considerations during the review process, and their timely input to ensure the integration of ethics review in acceptance recommendations and decisions.
- The ethics reviewers: Yonatan Bisk, Kevin Bretonnel Cohen, Francien Dechesne, Jack Hessel, Jin-Dong Kim, Anne Lauscher, Dave Lewis, Margot Mieskes, Xanda Schofield, Lyle Ungar, Jingbo Xia.
- The outstanding reviewers and action editors who were nominated by the senior area chairs for writing reviews that were particularly helpful in the decision making process. They are recognized by name later in this volume.

Experimenting with a new reviewing system on the large scale required by NAACL would not have been possible without the following people:

- Amanda Stent and Goran Glavaš, as ARR Editors-in-Chiefs, for their tireless work in support of the ARR December and January cycles.
- Graham Neubig, Dhruv Naik and Nils Dycke, as ARR Tech Team for these two cycles.
- Celeste Martinez Gomez, Melisa Bok, and Nadia L'Bahy, as OpenReview Tech Team.
- Elijah Rippeth for his help coordinating the special theme submissions.

The following committees helped shape the conference in countless ways:

- The best paper committee: Thamar Solorio (Chair), Isabelle Augenstein, Gemma Bel Enguix, Alona Fyshe, Shafiq Joty and Emily Prud'hommeaux who enthusiastically read and recommended papers for awards in a short time frame.
- The diversity & inclusion committee chaired by Diana Galván, Snigdha Chaturvedi and Yonatan Bisk, with advisors Pranav A and Luciana Benotti.
- The industry track chairs, Rashmi Gangadharaiah, Anastassia Loukina and Bonan Min, and advisors Owen Rambow and Yunyao Li.
- The tutorial chairs, Cecilia Alm, Miguel Ballesteros and Yulia Tsvetkov.
- The demonstration chairs, Hannaneh Hajishirzi, Qiang Ning and Avi Sil.
- The workshop chairs, Dan Goldwasser, Yunyao Li and Ashish Sabharwal.

- The student research workshop chairs, Daphne Ippolito, Liunian Harold Li, Maria Leonor Pacheco and advisors, Danqi Chen and Nianwen Xue.

- The publication chairs, Ryan Cotterell, Danilo Croce and Jordan Zhang.

- The reproducibility chairs, Niranjan Balasubramanian, Jesse Dodge, Annie Louis, Daniel Deutsch and Yash Kumar Lal.

- The sponsorship chair, Byron Wallace.

- The volunteer chair, Daniel Khashabi.

- The publicity chairs, Nanyun (Violet) Peng, Emily Sheng, Sameer Singh.

- The virtual infrastructure chairs, Deepak Ramachandran, Martín Villalba, and Rishita Anubhai.

- The website chairs, Ice Pasupat and Vered Shwartz for their exceptional reactivity and thorough checks of the conference schedule.

Finally, we would not have been able to organize this conference without the guidance, advice and cooperation of the following people:

- Damira Mrsic, Jernej Masnec, and Sol Rosenberg from Underline, who have been very prompt at answering all our questions and very helpful in setting up the virtual platform.

- Priscilla Rasmussen and Jenn Rachford who make all the logistics of the conference possible.

- Smaranda Muresan, Preslav Nakov, Aline Villavicencio, the Program co-Chairs of ACL 2022 who shared with us their materials and recent experience, and provided moral support.

- Anna Rumshisky, Thamar Solorio and Luke Zettlemoyer, as previous Program co-Chairs of NAACL, who answered many questions and provided invaluable guidance.

- TACL Editorial Assistant Cindy Robinson, and CL Editor-in-Chief Hwee Tou Ng for coordinating TACL and CL presentations with us.

- And last but not least, our General Chair, Dan Roth, whose guidance and support throughout the process were truly invaluable: his quiet strength, big picture thinking, and respect for all the parties involved were a soothing balm and an inspiration.

We hope you will enjoy the NAACL 2022 conference!

Marie-Catherine de Marneffe, Marine Carpuat and Ivan Vladimir Meza Ruiz
NAACL 2022 Program Committee Co-Chairs
June 2022

Organizing Committee

General Chair

Dan Roth, University of Pennsylvania and AWS AI Labs

Program Chairs

Marine Carpuat, University of Maryland
Marie-Catherine de Marneffe, Ohio State University
Ivan Vladimir Meza Ruiz, National Autonomous University of Mexico

Local Arrangements

Priscilla Rasmussen, ACL
Jennifer Rachford, ACL

Industry Track Chairs

Rashmi Gangadharaiah, Amazon
Anastassia Loukina, Grammarly
Bonan Min, Raytheon BBN Technologies

Advisors for the Industry Track

Owen Rambow, Stony Brook University
Yunyao Li, IBM Research

Tutorial Chairs

Cecilia Alm, Rochester Institute of Technology
Miguel Ballesteros, Amazon
Yulia Tsvetkov, University of Washington

Demonstration Chairs

Hannaneh Hajishirzi, University of Washington
Qiang Ning, Amazon
Avi Sil, IBM Research

Workshops Chairs

Dan Goldwasser, Purdue University
Yunyao Li, IBM Research
Ashish Sabharwal, Allen Institute for AI

Student Research Workshop Chairs

Daphne Ippolito, University of Pennsylvania
Liunian Harold Li, University of California Los Angeles
Maria Leonor Pacheco, Purdue University

Advisors for the Student Research Workshop

Danqi Chen, Princeton University
Nianwen Xue, Brandeis University

Publication Chairs

Ryan Cotterell, ETH Zürich
Danilo Croce, Tor Vergata University of Rome
Jordan Zhang,

Ethics Chairs

Kai-Wei Chang, University of California Los Angeles
Diyi Yang, Georgia Institute of Technology
Dirk Hovy, Bocconi University

Reproducibility Chairs

Niranjan Balasubramanian, Stony Brook University
Jesse Dodge, Allen Institute for AI
Annie Louis, Google
Daniel Deutsch, University of Pennsylvania
Yash Kumar Lal, Stony Brook University

Sponsorship Chair

Byron Wallace, Northeastern University

Diversity and Inclusion Chairs

Diana Galván, Tohoku University
Snigdha Chaturvedi, University of North Carolina Chapel Hill
Yonatan Bisk, Carnegie Mellon University

Advisory for the Diversity and Inclusion Committees

Pranav A, Dayta AI
Luciana Benotti, National University of Córdoba

Volunteers Chair

Daniel Khashabi, Allen Institute for AI

Publicity Chairs

Nanyun (Violet) Peng, Microsoft Research
Emily Sheng, University of North Carolina Chapel Hill
Sameer Singh, University of California Irvine

Virtual Infrastructure Chair

Deepak Ramachandran, Google
Martín Villalba, Saarland University
Rishita Anubhai, Amazon

Website Chairs

Ice Pasupat, Google
Vered Shwartz, University of British Columbia

Program Committee

Computational Social Science and Cultural Analytics

Svitlana Volkova, Pacific Northwest National Laboratory
David Bamman, University of California Berkeley

Dialogue and Interactive systems

Michel Galley, Microsoft
Kallirroi Georgila, University of Southern California
Nina Dethlefs, University of Hull
Heriberto Cuayáhuitl, University of Lincoln

Discourse and Pragmatics

Viviane Moreira, Universidade Federal do Rio Grande do Sul
Nafise Moosavi, University of Sheffield

Ethics Bias and Fairness

Vinodkumar Prabhakaran, Google
Svetlana Kiritchenko, National Research Council Canada

Efficient methods in NLP

Alexandra Luccioni, Hugging Face
Roy Schwartz, Hebrew University Hebrew University of Jerusalem

Language Generation

Michael White, Ohio State University
Snigdha Chaturvedi, Department of Computer Science University of North Carolina Chapel Hill
Shashi Narayan, Google

Information Extraction

Muhao Chen, University of Southern California
Timothy Miller, Harvard University
Deepak Ramachandran, Google
Ruihong Huang, Texas A&M University

Information Retrieval and Text Mining

Sophia Ananiadou, University of Manchester
Luca Soldaini, Allen Institute for Artificial Intelligence

Interpretability and Analysis of Models for NLP

Sebastian Gehrmann, Google Research
Sameer Singh, University of California Irvine

Language Grounding to Vision Robotics and Beyond

Parisa Kordjamshidi, Michigan State University
Peter Anderson, Google

Language Resources and Evaluation

Annemarie Friedrich, Bosch Center for Artificial Intelligence
Sebastian Schuster, New York University
Pradeep Dasigi, Allen Institute for Artificial Intelligence

Linguistic Theories Cognitive Modeling and Psycholinguistics

Allyson Ettinger, University of Chicago
Raquel Fernández, University of Amsterdam

Machine Learning for NLP - Classification and Structured Prediction Models

He He, New York University
Wei Xu, Georgia Institute of Technology

Machine Learning for NLP - Language Modeling and Sequence to Sequence Models

Colin Raffel, Hugging Face
Miguel Ballesteros, Amazon

Machine Translation

Kevin Duh, Johns Hopkins University
Gholamreza Haffari, Monash University
Rachel Bawden, Inria

Multilinguality

Dan Garrette, Google Research
Avirup Sil, International Business Machines

NLP Applications

Vukosi Marivate, University of Pretoria
Helena Gomez Adorno, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas - UNAM
Wei Lu, Singapore University of Technology and Design
Dan Goldwasser, Purdue University

Phonology Morphology and Word Segmentation

Micha Elsner, Ohio State University
Reut Tsarfaty, Bar-Ilan University Technion

Question Answering

Mihaela Bornea, IBM International Business Machines
Jordan Boyd-Graber, University of Maryland College Park
Siva Reddy, Mila McGill University

Semantics - Lexical Semantics

Adina Williams, Facebook AI Research (Meta Platforms Inc.)
Mohammad Taher Pilehvar, Tehran Institute for Advanced Studies

Semantics - Sentence-level Semantics and Textual Inference

Xiaodan Zhu, Queen's University
Eduardo Blanco, Arizona State University
Rachel Rudinger, University of Maryland College Park

Sentiment Analysis and Stylistic Analysis

Davide Buscaldi, Ecole polytechnique
Sara Rosenthal, International Business Machines

Speech

Adam Stiff, Infoscitex Corporation
Rolando Coto Solano, Dartmouth College

Summarization

Greg Durrett, University of Texas Austin
Jackie Cheung, Microsoft

Syntax - Tagging Chunking and Parsing

Marie Candito, Université de Paris
Lilja Øvrelid, Dept. of Informatics University of Oslo

Special Theme

Jeffrey Bigham, Carnegie Mellon University

Action Editors

Zeljko Agic, Md Shad Akhtar, Malihe Alikhani, Cecilia Alm, Mark Anderson, Jacob Andreas, Xiang Ao, Marianna Apidianaki, Yuki Arase, Mikel Artetxe, Ehsaneddin Asgari, Giuseppe Attardi, Niranjan Balasubramanian, Timothy Baldwin, Miguel Ballesteros, Mohamad Hardyman Barawi, Jeremy Barnes, Loïc Barrault, Roberto Basili, Ali Basirat, Jasmijn Bastings, Daniel Beck, Iz Beltagy, Luciana Benotti, Steven Bethard, Chandra Bhagavatula, Alexandra Birch, Yonatan Bisk, Danushka Bollegala, Florian Boudin, Leonid Boytsov, Chris Brew, Elena Cabrio, Aoife Cahill, Andrew Caines, Ruken Cakici, Jose Camacho-Collados, Yanan Cao, Ziqiang Cao, Cornelia Caragea, Paula Carvalho, Andrew Cattle, Daniel Cer, Muthukumar Chandrasekaran, Angel Chang, Kai-Wei Chang, Boxing Chen, Danqi Chen, Kuan-Yu Chen, Lei Chen, Colin Cherry, Hai

Leong Chieu, Luis Chiruzzo, Eunsol Choi, Jinho Choi, Prafulla Kumar Choubey, Khalid Choukri, Oana Cocarascu, John Conroy, Caio Corro, Marta Costa-jussà, Aron Culotta, Raj Dabre, Daniel Dakota, Dipanjan Das, Johannes Daxenberger, Pascal Denis, Barry Devereux, Georgiana Dinu, Jesse Dodge, Li Dong, Eduard Dragut, Kevin Duh, Miryam de Lhoneux, Marie-Catherine de Marneffe, Liat Ein-Dor, Michael Elhadad, Allyson Ettinger, Angela Fan, Anna Feldman, Naomi Feldman, Xiaocheng Feng, Yang Feng, Yansong Feng, Francis Ferraro, Elisabetta Fersini, Simone Filice, Mark Fishel, Pascale Fung, Matthias Gallé, Zhe Gan, Yang Gao, Alborz Geramifard, Debanjan Ghosh, Goran Glavaš, Kyle Gorman, Jiatao Gu, Qing Gu, Honglei Guo, Hongyu Guo, Qipeng Guo, Nizar Habash, Ivan Habernal, Christian Hardmeier, Yulan He, Zhongjun He, Daniel Hershcovich, Julia Hockenmaier, Enamul Hoque, Baotian Hu, Shujian Huang, Xuanjing Huang, Ozan Irsoy, Srini Iyer, Cassandra Jacobs, Kokil Jaidka, Hyeju Jang, Yangfeng Ji, Preethi Jyothi, Sarvnaz Karimi, Shubhra Kanti Karmaker, Daisuke Kawahara, Daniel Khashabi, Jin-Dong Kim, Seokhwan Kim, Taeuk Kim, Ekaterina Kochmar, Grzegorz Kondrak, Amrith Krishna, Udo Kruschwitz, Marco Kuhlmann, Sumeet Kumar, Jonathan Kummerfeld, Wai Lam, Zhenzhong Lan, Mark Last, Hady Lauw, Carolin Lawrence, John Lawrence, Alessandro Lenci, Lori Levin, Mike Lewis, Patrick Lewis, Jing Li, Junhui Li, Juntao Li, Liangyou Li, Piji Li, Sujian Li, Wenjie Li, Maria Liakata, Constantine Lignos, Dekang Lin, Marco Lippi, Pengfei Liu, Qun Liu, Yang Liu, Zhiyuan Liu, Anh Tuan Luu, Wei-Yun Ma, Craig MacDonald, Andrea Madotto, Navonil Majumder, Prodromos Malakasiotis, Igor Malioutov, Eugenio Martinez-Camara, Bruno Martins, Yuji Matsumoto, . Mausam, David McClosky, Mahnoosh Mehrabani, Margot Mieskes, Makoto Miwa, Daichi Mochihashi, Mohamed Morchid, Antonio Moreno-Ortiz, David Mortensen, Lili Mou, Philippe Muller, Kenton Murray, Nona Naderi, Courtney Napoles, Shashi Narayan, Roberto Navigli, Mark-Jan Nederhof, Vincent Ng, Dat Quoc Nguyen, Thien Nguyen, Jan Niehues, Qiang Ning, Maciej Ogrodniczuk, Alice Oh, Naoaki Okazaki, Manabu Okumura, Matan Orbach, Miles Osborne, Jessica Ouyang, Ankur Parikh, Joonsuk Park, Seong-Bae Park, Yannick Parmentier, Tommaso Pasini, Rebecca Passonneau, Viviana Patti, Nanyun Peng, Laura Perez-Beltrachini, Sandro Pezzelle, Juan Pino, Emily Pitler, Barbara Plank, Edoardo Ponti, Simone Ponzetto, Kashyap Popat, Maja Popovic, Soujanya Poria, Vinodkumar Prabhakaran, Daniel Preotiuc-Pietro, Emily Prud'hommeaux, Tieyun Qian, Xipeng Qiu, Xiaojun Quan, Alessandro Raganato, Ganesh Ramakrishnan, Siva Reddy, Ines Rehbein, Roi Reichart, Xiang Ren, Yafeng Ren, Sebastian Riedel, Joseph Roux, Alla Rozovskaya, Attapol Rutherford, Diarmuid Séaghdha, Alexandre Salle, Maarten Sap, Hinrich Schütze, Timo Schick, Nathan Schneider, H. Schwartz, Lane Schwartz, Minjoon Seo, Bei Shi, Tianze Shi, Lei Shu, Melanie Siegel, Gabriel Skantze, Kevin Small, Yangqiu Song, Vivek Srikumar, Shashank Srivastava, Efstathios Stamatatos, Gabriel Stanovsky, Amanda Stent, Karl Stratos, Emma Strubell, Sara Stymne, Saku Sugawara, Jun Suzuki, Dima Taji, Duyu Tang, Harish Tayyar Madabushi, Paolo Torroni, Trang Tran, Chen-Tse Tsai, Jun'ichi Tsujii, Kewei Tu, Stefan Ultes, Olga Vechtomova, Giulia Venturi, Suzan Verberne, Yannick Versley, David Vilares, Thuy Vu, Ivan Vulić, Yogarshi Vyas, Byron Wallace, Xiaojun Wan, Longyue Wang, Shuai Wang, Xin Wang, Zhiguang Wang, Leo Wanner, Zeerak Waseem, Shinji Watanabe, Zhongyu Wei, Rodrigo Wilkens, Alina Wróblewska, Lijun Wu, Tong Xiao, Deyi Xiong, Hainan Xu, Rui Yan, Min Yang, Jin-Ge Yao, Wenlin Yao, Wenpeng Yin, Koichiro Yoshino, Jianfei Yu, Kai Yu, Mo Yu, François Yvon, Marcos Zampieri, Marcely Zanon Boito, Fabio Massimo Zanzotto, Amir Zeldes, Luke Zettlemoyer, Justine Zhang, Weinan Zhang, Xiangliang Zhang, Xingxing Zhang, Yi Zhang, Yue Zhang, Zhe Zhang, Michael Zock

Reviewers

Robert Östling, Ahmet Üstün, Erion Çano, Blaž Škrlj, Micheal Abaho, Ahmed Abdelali, Muhammad Abdul-Mageed, Omri Abend, Karl Aberer, Lasha Abzianidze, David Adelani, Somak Aditya, Noëmi Aepli, Divyansh Agarwal, Sanchit Agarwal, Rodrigo Agerri, Milan Aggarwal, Piush Aggarwal, Manex Agirrezabal, Guy Aglionby, Ameeta Agrawal, Priyanka Agrawal, Sweta Agrawal, Roee Aharoni, Wasi Ahmad, Benyamin Ahmadnia, Murtadha Ahmed, Aman Ahuja, Kabir Ahuja, Chunhui Ai, Joshua Ainslie, Akiko Aizawa, Reina Akama, VSDS Mahesh Akavarapu, Alan Akbik, Syed Sarfaraz Akhtar, Nader Akoury, Arjun Akula, Ekin Akyurek, Hend Al-Khalifa, Nora Al-Twairesh, Amal Alabdulkarim, Ozge Alacam, Firoj Alam, Georgios Alexandridis, Dmitriy Alexandrov, David Alfter, Bashar Alhafni, Raquel Alhama, Tariq Alhindi, Hassam Alhuzali, Mohammad Aliannejadi, Tamer Alkhouli, Emily Allaway, Manar Alohaly, Héctor Alonso, Miguel Alonso, Sawsan Alqahtani, Milad Alshomary, Sophia Althammer, Fernando Alva-Manchego, Rami Aly, Aixiu An, Jisun An, Nicholas Andrews, Gabor Angeli, Diego Antognini, Jean-Yves Antoine, Kaveri Anuranjana, Emilia Apostolova, Negar Arabzadeh, Eiji Aramaki, Arturo Argueta, Mozhdeh Ariannezhad, Naveen Arivazhagan, Ravneet Arora, Ignacio Arroyo-Fernández, Katya Artemova, Yoav Artzi, Akari Asai, Elliott Ash, Zhenisbek Assylbekov, Duygu Ataman, Ben Athiwaratkun, Giuseppe Attanasio, Isabelle Augenstein, Michael Auli, Amir Avestimehr, Eleftherios Avramidis, Parul Awasthy, Hosein Azarbonyad, Wilker Aziz, Matthias Aßenmacher, Bogdan Babych, Sanghwan Bae, Deblin Bagchi, Ebrahim Bagheri, Dzmitry Bahdanau, Ashutosh Baheti, Vikas Bahirwani, Fan Bai, He Bai, Long Bai, Yu Bai, JinYeong Bak, Vidhisha Balachandran, Mithun Balakrishna, Tyler Baldwin, Nicolae Banari, Juan Banda, Pratyay Banerjee, Jeesoo Bang, Hritik Bansal, Sameer Bansal, Hangbo Bao, Jianzhu Bao, Junwei Bao, Yu Bao, Zuyi Bao, Ankur Bapna, Roy Bar-Haim, Edoardo Barba, Denilson Barbosa, Ken Barker, Leslie Barrett, James Barry, Max Bartolo, Marco Basaldella, Pierpaolo Basile, Valerio Basile, Mohaddeseh Bastan, Daniel Bauer, Timo Baumann, Kathy Baxter, Tilman Beck, Lee Becker, Dorothee Beermann, Lisa Beinborn, Ahmad Beirami, Giannis Bekoulis, Gemma Bel-Enguix, Eric Bell, Meriem Beloucif, Eyal Ben-David, Yassine Benajiba, Michael Bendersky, Luisa Bentivogli, Adrian Benton, Jonathan Berant, Alexandre Berard, Gábor Berend, Taylor Berg-Kirkpatrick, Maria Berger, Toms Bergmanis, Rafael Berlanga, Timothée Bernard, Delphine Bernhard, Dario Bertero, Aditya Bhargava, Suma Bhat, Sumit Bhatia, Kasturi Bhattacharjee, Pushpak Bhattacharyya, Satwik Bhattamishra, Shruti Bhosale, Rajarshi Bhowmik, Victoria Bi, Laura Biester, Philippe Blache, Philip Blair, Terra Blevins, Rexhina Blloshmi, Jelke Bloem, Michael Bloodgood, Sravan Babu Bodapati, Ben Bogin, Nikolay Bogoychev, Bernd Bohnet, Ondrej Bojar, Gemma Boleda, Danushka Bollegala, Valeriia Bolotova, Daniele Bonadiman, Logan Born, Piyush Borole, Emanuela Boros, Tulika Bose, Antoine Bosselut, Robert Bossy, Kaj Bostrom, Nadjet Bouayad-Agha, Zied Bouraoui, Samuel Bowman, Johan Boye, Kristy Boyer, Faeze Brahman, Antonio Branco, Stephanie Brandl, Arthur Brazinskas, Jonathan Brennan, Chris Brew, Eleftheria Briakou, Thomas Brochhagen, Samuel Broscheit, Thomas Brovelli, Pawel Budzianowski, Sven Buechel, Emanuele Bugliarello, Joan Byamugisha, Bill Byrne, Avi Caciularu, Samuel Cahyawijaya, Deng Cai, Han Cai, Hengyi Cai, Pengshan Cai, Yi Cai, Agostina Calabrese, Iacer Calixto, Jesus Calvillo, Jose Camacho-Collados, Erik Cambria, Oana-Maria Camburu, Giovanni Campagna, Leonardo Campillos-Llanos, Niccolò Campolungo, Daniel Campos, Jon Ander Campos, Burcu Can, Jie Cao, Juan Cao, Kris Cao, Qingqing Cao, Qingxing Cao, Ruisheng Cao, Shuyang Cao, Steven Cao, Yixin Cao, Yu Cao, Yuan Cao, Yunbo Cao, Annalina Caputo, Dallas Card, Ronald Cardenas, Rémi Cardon, Luigi Caro, Lucien Carroll, Danilo Carvalho, Tommaso Caselli, Justine Cassell, Vittorio Castelli, Giuseppe Castellucci, Paulo Cavalin, Arun Tejasvi Chaganty, Soumen Chakrabarti, Abhisek Chakrabarty, Tuhin Chakrabarty, Sunandan Chakraborty, Bharathi Raja Chakravarthi, Ilias Chalkidis, Alvin Chan, Hou Pong Chan, Zhangming Chan, Khyathi Chandu, Serina Chang, Xuankai Chang, Yung-Chun Chang, Soravit Changpinyo, WenHan Chao, Shubham Chatterjee, Akshay Chaturvedi, Konstantinos Chatzisavvas, Aditi Chaudhary, Kushal Chawla, Wanxiang Che, Gullal Singh Cheema, Saneem Chemmengath, Bei Chen, Bo Chen, Cen Chen, Chenhua Chen, Chung-Chi Chen, Danqi Chen, Daoyuan Chen, Feilong Chen, Guangyong Chen, Guanhua Chen, Guanyi Chen, Hanjie Chen, Hong Chen, Howard Chen, Jiaze Chen, Jifan Chen, John Chen, Kezhen Chen, Lei Chen, Lin Chen, Lu Chen, Luoxin Chen, Meng Chen, Mingda Chen, Pei Chen, Qian Chen, Qianglong Chen, Qingcai Chen, Tianlang Chen, Tongfei Chen, Wang Chen, Wei-Fan Chen, Wenhu Chen, Wenliang Chen, Wenqing Chen, Xilun Chen, Xinchi Chen, Xiuying Chen, Yen-Chun Chen, Yubo Chen, Yue Chen, Yufeng Chen, Yulong Chen, Yun Chen, Zhi Chen, Zhihong Chen, Zhuang Chen, Zhumin Chen, Zhuohao Chen, Liying Cheng, Lu Cheng, Minhao Cheng, Pengxiang Cheng, Pengyu Cheng, Shanbo Cheng, Weiwei Cheng, Yong Cheng, Zifeng Cheng, Emmanuele Chersoni, Ethan Chi, Ta-Chung Chi, Zewen Chi, Yew Ken Chia, David Chiang, Ting-Rui Chiang, Patricia Chiril, Nadezhda Chirkova, Francisco Chiyah-Garcia, Hyunsoo Cho, Jaemin Cho, Kyunghyun Cho, Sangwoo Cho, Won Ik Cho, Eunsol Choi, Jinho Choi, Jonghyun Choi, Seungtaek Choi, Shamil Chollampatt, Jaegul Choo, Leshem Choshen, Prafulla Kumar Choubey, Jishnu Chowdhury, Md Faisal Mahbub Chowdhury, Somnath Chowdhury, Christos Christodoulopoulos, Fenia Christopoulou, Alexandra Chronopoulou, Chenhui Chu, Christopher Chu, Tat-seng Chua, Jin-Woo Chung, Kenneth Church, Abu Nowshed Chy, Volkan Cirik, Christopher Clark, Elizabeth Clark, Christopher Clarke, Miruna Clinciu, Trevor Cohen, Jeremy Cole, Marcus Collins, Pierre Colombo, Kathryn Conger, Simone Conia, Mathieu Constant, Noah Constant, Danish Contractor, Robin Cooper, Anna Corazza, Luciano Corro, Josep Crego, Danilo Croce, Fabien Cromieres, Lei Cui, Leyang Cui, Shaobo Cui, Xia Cui, Yiming Cui, Anna Currey, Tonya Custis, Luis Fernando D'Haro, Jennifer D'Souza, Hervé Déjean, Giovanni Da San Martino, Raj Dabre, Deborah Dahl, Hongliang Dai, Wenliang Dai, Xiang Dai, Xinyu Dai, Yinpei Dai, Zhuyun Dai, Siddharth Dalmia, Sandipan Dandapat, Marina Danilevsky, Verna Dankers, Anubrata Das, Rajarshi Das, Sarthak Dash, Vidas Daudaravicius, Sam Davidson, Brian Davis, Ernest Davis, Heidar Davoudi, Steve DeNeefe, Alok Debnath, Marco Del Tredici, Louise Deleger, Agustín Delgado, David Demeter, Xiang Deng, Yang Deng, Yuntian Deng, Zhenyun Deng, Zhongfen Deng, Tejaswini Deoskar, Jan Deriu, Franck Dernoncourt, Ameet Deshpande, Roberto Dessi, Tim Dettmers, Sunipa Dev, Jwala Dhamala, Maria Pia Di Buono, Shizhe Diao, Gaël Dias, Emily Dinan, Chenchen Ding, Haibo Ding, Kaize Ding, Liang Ding, Ning Ding, Shuoyang Ding, Long Doan, Charles Dognin, Miguel Domingo, Lucia Donatelli, Domenic Donato, Haoyu Dong, Qian Dong, Yue Dong, Bonaventure F. P. Dossou, Longxu Dou, Zi-Yi Dou, Doug Downey, Mark Dras, Markus Dreyer, Rotem Dror, Aleksandr Drozd, Andrew Drozdov, Jingfei Du, Jinhua Du, Li Du, Mengnan Du, Pan Du, Wanyu Du, Xinya Du, Yupei Du, Xiangyu Duan, Kumar Dubey, Pablo Duboue, Shiran Dudy, Philipp Dufter, Jonathan Dunn, Benjamin Durme, Ritam Dutt, Gaël de Chalendar, Gustavo de Rosa, Flor Miriam del Arco, Dylan Ebert, Abteen Ebrahimi, Aleksandra Edwards, Steffen Eger, Markus Egg, Yo Ehara, Vladimir Eidelman, Bryan Eikema, Jacob Eisenstein, Adam Ek, Asif Ekbal, Aparna Elangovan, Yanai Elazar, Heba Elfardy, Michael Elhadad, AbdelRahim Elmadany, Micha Elsner, Denis Emelin, Guy Emerson, Akiko Eriguchi, Liana Ermakova, Ori Ernst, Patrick Ernst, Carlos Escolano, Arash Eshghi, Ramy Eskander, Cristina España-Bonet, Luis Espinosa-Anke, Kawin Ethayarajh, Kilian Evang, Richard Evans, Michael Färber, Alexander Fabbri, Guglielmo Faggioli, Martin Fajčík, Neele Falk, Tobias Falke, Chuang Fan, Feifan Fan, Lu Fan, Anjie Fang, Wei Fang, Yihao Fang, Yuwei Fang, Hossein Fani, Amir Feder, Hao Fei, Nils Feldhus, Mariano Felice, Jiazhan Feng, Junlan Feng, Shi Feng, Shi Feng, Shikun Feng, Xiachong Feng, Yansong Feng, Emmanouil Fergadiotis, James Ferguson, Daniel Fernández-González, Patrick Fernandes, Elisa Ferracane, Javier Ferrando, Francis Ferraro, Thiago Ferreira, Olivier Ferret, Besnik Fetahu, Anjalie Field, Alejandro Figueroa, Katja Filippova, Catherine Finegan-Dollak, Nicholas FitzGerald, Margaret Fleck, Lucie Flekova, Antske Fokkens, Marina Fomicheva, José Fonollosa, Tommaso Fornaciari, Paula Fortuna, Eric Fosler-Lussier, George Foster, Mary Ellen Foster, James Foulds, Thomas François, Anette Frank, Stella Frank, Alexander Fraser, Kathleen Fraser, Marjorie Freedman, Dayne Freitag, Markus Freitag, Andre Freitas, Lea Frermann, Daniel Fried, Guohong Fu, Jie Fu, Jinlan Fu, Peng Fu, Tsu-Jui Fu, Xingyu Fu, Xiyan Fu, Yao Fu, Yoshinari Fujinuma, Kotaro Funakoshi, Adam Funk, Jana Götze, Carlos Gómez-Rodríguez, Matteo Gabburo, Saadia Gabriel, David Gaddy, Marco Gaido, Andrea Galassi, Mark Gales, Leilei Gan, Yujian Gan, Zhe Gan, Kuzman Ganchev, Sudeep Gandhe, Ashwinkumar Ganesan, Balaji Ganesan, Debasis Ganguly, William Gantt, Ge Gao, Jie Gao, Jun Gao, Shen Gao, Tianyu Gao, Yifan Gao, Yingbo Gao, Matt Gardner, Dinesh Garg, Muskan Garg, Sarthak Garg, Siddhant Garg, Albert Gatt, Manas Gaur, Dipesh Gautam, Shen Ge, Yubin Ge, Ruiying Geng, Ariel Gera, Mor Geva, Demian Ghalandari, Sepideh Ghanavati, Sarik Ghazarian, Sarik Ghazarian, Marjan Ghazvininejad, Mozhdeh Gheini, Ahmed Ghoneim, Deepanway Ghosal, Debanjan Ghosh, Sourav Ghosh, Sucheta Ghosh, Daniel Gildea, Salvatore Giorgi, Voula Giouli, Adrià Gispert, Mario Giulianelli, Michael Glass, Goran Glavaš, Alfio Gliozzo, Ameya Godbole, Vaibhava Goel, Nazli Goharian, Tejas Gokhale, Sujatha Das Gollapalli, Marcos Goncalves, Lovedeep Gondara, Heng Gong, Hongyu Gong, Samuel

Gonzalez-Lopez, Jeffrey Good, Rob Goot, Karthik Gopalakrishnan, Vera Gor, Philip Gorinski, Isao Goto, Cyril Goutte, Kartik Goyal, Naman Goyal, Pawan Goyal, Tanya Goyal, Mario Graff, Christophe Gravier, Yulia Grishina, Milan Gritta, Loïc Grobol, Karol Grzegorczyk, Jia-Chen Gu, Jing Gu, Jian Guan, Saiping Guan, Shuo Guan, Yi Guan, Marco Guerini, Lin Gui, Vincent Guigue, Liane Guillou, Camille Guinaudeau, Kalpa Gunaratna, Chulaka Gunasekara, James Gung, Tunga Gungor, Jiang Guo, Jiaqi Guo, Junliang Guo, Ruocheng Guo, Yinpeng Guo, Zhijiang Guo, Abhirut Gupta, Prakhar Gupta, Rahul Gupta, Shashank Gupta, Sparsh Gupta, Vivek Gupta, Izzeddin Gur, Suchin Gururangan, Joakim Gustafson, Ximena Gutierrez-Vasques, Ido Guy, Ali Hürriyetoğlu, Jung-Woo Ha, Nizar Habash, Ivan Habernal, Christian Hadiwinoto, Nafaa Haffar, Matthias Hagen, Michael Hahn, Zhen Hai, Huda Hakami, David Hall, Michael Hammond, Jialong Han, Jiawei Han, Lifeng Han, Namgi Han, Rujun Han, Seungju Han, Shi Han, Wenjuan Han, Xianpei Han, Xudong Han, Abram Handler, Greg Hanneman, Jie Hao, Momchil Hardalov, Hardy Hardy, Mareike Hartmann, Thomas Hartvigsen, Peter Hase, Chikara Hashimoto, Hany Hassan, Nabil Hathout, Bradley Hauer, Hiroaki Hayashi, Katsuhiko Hayashi, Yoshihiko Hayashi, Shirley Hayati, Devamanyu Hazarika, Timothy Hazen, Ben He, Junxian He, Liang He, Liang He, Luheng He, Shizhu He, Taiqi He, Wanwei He, Xuanli He, Yifan He, Youbiao He, Behnam Hedayatnia, Jindřich Helcl, Matthew Henderson, Leonhard Hennig, Nico Herbig, Jonathan Herzig, Jack Hessel, John Hewitt, Rem Hida, Djoerd Hiemstra, Ryuichiro Higashinaka, Xavier Hinaut, Tsutomu Hirao, Tatsuya Hiraoka, Nils Hjortnaes, Cuong Hoang, Hieu Hoang, Johannes Hoffart, Valentin Hofmann, Chris Hokamp, Nora Hollenstein, Ari Holtzman, Takeshi Homma, Ukyo Honda, Xudong Hong, Mark Hopkins, Andrea Horbach, Md Mosharaf Hossain, Saghar Hosseini, Feng Hou, Lei Hou, Yutai Hou, Dirk Hovy, David Howcroft, Estevam Hruschka, Shu-Kai Hsieh, Benjamin Hsu, Chao-Chun Hsu, Chun-Nan Hsu, I-Hung Hsu, Wei-Ning Hsu, Phu-Mon Htut, Chi Hu, Hexiang Hu, Jennifer Hu, Jinyi Hu, Minghao Hu, Pengwei Hu, Po Hu, Renfen Hu, Wei Hu, Zhe Hu, Ziniu Hu, Xinyu Hua, Chao-Wei Huang, Danqing Huang, He Huang, Hen-Hsen Huang, Heyan Huang, James Y. Huang, Jiaji Huang, Jie Huang, Kaiyu Huang, Kuan-Hao Huang, Kung-Hsiang Huang, Minlie Huang, Quzhe Huang, Songfang Huang, Xiaolei Huang, Xinting Huang, Zhongqiang Huang, Patrick Huber, Binyuan Hui, Kai Hui, Chia-Chien Hung, Dieuwke Hupkes, Ben Hutchinson, Tin Huynh, Jena Hwang, Sung Ju Hwang, Ignacio Iacobacci, Timour Igamberdiev, Oana Ignat, Alvin Ii, Ryu Iida, Taichi Iki, Gabriel Ilharco, Filip Ilievski, Nikolai Ilinykh, Irina Illina, Dmitry Ilvovsky, Mert Inan, Sathish Reddy Indurthi, Go Inoue, Koji Inoue, Naoya Inoue, Radu Ionescu, Daphne Ippolito, Hitoshi Isahara, Tatsuya Ishigaki, Neslihan Iskender, Tunazzina Islam, Hayate Iso, Takumi Ito, Tomoya Iwakura, Kenichi Iwatsuki, Aaron Jaech, Sarthak Jain, Sujay Kumar Jauhar, Tommi Jauhiainen, Pratik Jayarao, Sébastien Jean, Sungho Jeon, Kevin Jesse, Harsh Jhamtani, Houye Ji, Shouling Ji, Zongcheng Ji, Chen Jia, Ran Jia, Robin Jia, Ruoxi Jia, Yuxiang Jia, Ping Jian, Daxin Jiang, Hongfei Jiang, Huixing Jiang, Jing Jiang, Meng Jiang, Nan Jiang, Nanjiang Jiang, Shaojie Jiang, Wenbin Jiang, Xin Jiang, Zhengbao Jiang, Zhihua Jiang, Zhuolin Jiang, Zhuoren Jiang, Zhuoxuan Jiang, Wenxiang Jiao, Allan Jie, Di Jin, Hailong Jin, Lifeng Jin, Peng Jin, Xiaolong Jin, Yiping Jin, Xu Jinan, Baoyu Jing, Yohan Jo, Richard Johansson, Shailza Jolly, Erik Jones, Kenneth Joseph, Sachindra Joshi, Dhanya Jothimani, Xincheng Ju, Jaap Jumelet, Heewoo Jun, Prathyusha Jwalapuram, Jad Kabbara, Kazuma Kadowaki, Tomoyuki Kajiwara, Mihir Kale, Oren Kalinsky, Aikaterini-Lida Kalouli, Hidetaka Kamigaito, Jaap Kamps, Min-Yen Kan, Hiroshi Kanayama, Masahiro Kaneko, Minki Kang, Yoshinobu Kano, Diptesh Kanojia, Evangelos Kanoulas, Georgi Karadzhov, Pinar Karagoz, Siddharth Karamcheti, Mladen Karan, Younes Karimi, Börje Karlsson, Sanjeev Karn, Jungo Kasai, Omid Kashefi, Yosuke Kashiwagi, Zdeněk Kasner, Nora Kassner, Divyansh Kaushik, Pride Kavumba, Hideto Kazawa, Amirhossein Kazemnejad, Abe Kazemzadeh, Pei Ke, Zixuan Ke, Brendan Kennedy, Casey Kennington, Tom Kenter, Daniel Kershaw, Santosh Kesiraju, Madian Khabsa, Salam Khalifa, Jawad Khan, Dinesh Khandelwal, Urvashi Khandelwal, Simran Khanuja, Mitesh Khapra, Aparna Khare, Khalid Khatib, Mikhail Khodak, Tushar Khot, Johannes Kiesel, Dong-Jin Kim, Gene Kim, Gunhee Kim, Gyuwan Kim, Hyounghun Kim, Hyunwoo Kim, Jaehyung Kim, Jihyuk Kim, Joo-Kyung Kim, Juyong Kim, Kang-Min Kim, Sungdong Kim, Yeachan Kim, Yoon Kim, Tracy King, Christo Kirov, Nikita Kitaev, Hirokazu Kiyomaru, Shun Kiyono, Ayal Klein, Bennett Kleinberg, Mateusz Klimaszewski, Rebecca Knowles, Hideo

Kobayashi, Elena Kochkina, Jordan Kodner, Philipp Koehn, Svetla Koeva, Mare Koit, Noriyuki Kojima, Rik Koncel-Kedziorski, Cunliang Kong, Lingkai Kong, Miloslav Konopík, Moshe Koppel, Yuta Koreeda, Mandy Korpusik, Katsunori Kotani, Fajri Koto, Venelin Kovatchev, Sebastian Krause, Elisa Kreiss, Simon Krek, Ralf Krestel, Florian Kreyssig, Kalpesh Krishna, Kundan Krishna, Balaji Krishnamurthy, Nikhil Krishnaswamy, Reno Kriz, Canasai Kruengkrai, Udo Kruschwitz, Germàn Kruszewski, Lun-Wei Ku, Taras Kucherenko, Mayank Kulkarni, Sayali Kulkarni, Vivek Kulkarni, Artur Kulmizev, Devang Kulshreshtha, Ashutosh Kumar, Ayush Kumar, Dhruv Kumar, Sachin Kumar, Sawan Kumar, Shankar Kumar, Varun Kumar, Anoop Kunchukuttan, Souvik Kundu, Florian Kunneman, Jenny Kunz, Tatsuki Kuribayashi, Shuhei Kurita, Kemal Kurniawan, Robin Kurtz, Andrey Kutuzov, Matthieu Labeau, Oier Lacalle, Faisal Ladhak, Huiyuan Lai, Viet Lai, Yuxuan Lai, Vasudev Lal, Yash Kumar Lal, Divesh Lala, John Lalor, Tsz Kin Lam, Wai Lam, Matthew Lamm, Gerasimos Lampouras, Man Lan, Wuwei Lan, Yunshi Lan, Ni Lao, Guy Lapalme, Mirella Lapata, Gabriella Lapesa, Stefan Larson, Alex Lascarides, Md Tahmid Rahman Laskar, Anne Lauscher, Alberto Lavelli, Dawn Lawrie, Henry Le, Phong Le, Matthew Lease, Andrew Lee, Chia-Hsuan Lee, Dongkyu Lee, Dongyub Lee, Fei-Tzin Lee, Hung-yi Lee, Hyunju Lee, Ji-Ung Lee, Joosung Lee, Kenton Lee, Nayeon Lee, Sang-Woo Lee, Seolhwa Lee, Young-Suk Lee, Artuur Leeuwenberg, Tao Lei, Wenqiang Lei, Jochen Leidner, Alessandro Lenci, Yichong Leng, Piyawat Lertvittayakumjorn, Brian Lester, Guy Lev, Gina-Anne Levow, Sharon Levy, Ashley Lewis, Martha Lewis, Bai Li, Bo Li, Bowen Li, Bryan Li, Chen Li, Chenliang Li, Chenliang Li, Chunyuan Li, Dianqi Li, Dingcheng Li, Dongfang Li, Haizhou Li, Haonan Li, Haoran Li, Hongyu Li, Huayang Li, Irene Li, Jialu Li, Jinchao Li, Jing Li, Jingjing Li, Jingye Li, Jiwei Li, Juanzi Li, Juncheng Li, Liangyou Li, Lin Li, Manling Li, Miao Li, Minglei Li, Peifeng Li, Penh Li, Ruifan Li, Ruizhe Li, Shang-Wen Li, Shaohua Li, Sheng Li, Shuangyin Li, Shuyang Li, Tao Li, Xiang Li, Xiangci Li, Xiaonan Li, Xin Li, Xintong Li, Xiujun Li, Yanran Li, Yanzeng Li, Yaoyiran Li, Ying Li, Yingjie Li, Yingya Li, Yinqiao Li, Yitong Li, Yiyuan Li, Yuan-Fang Li, Yuliang Li, Zhen Li, Zheng Li, Zhenghua Li, Zhongyang Li, Zhoujun Li, Zichao Li, Zongxi Li, Zuchao Li, Bin Liang, Chao-Chun Liang, Chen Liang, Paul Pu Liang, Shining Liang, Yunlong Liang, Zhicheng Liang, Bei Liao, Lizi Liao, Jindřich Libovický, Chaya Liebeskind, Wang Lijie, Gilbert Lim, Kwan Lim, Bill Yuchen Lin, Guan-Ting Lin, Hongfei Lin, Junyang Lin, Kevin Lin, Lucy Lin, Peiqin Lin, Thomas Lin, Tony Lin, Weizhe Lin, Xi Lin, Xiang Lin, Yankai Lin, Ying Lin, Zhaojiang Lin, Zheng Lin, Zi Lin, Tal Linzen, Pierre Lison, Robert Litschko, Patrick Littell, Marina Litvak, Alisa Liu, Bin Liu, Bing Liu, Chen Liu, Chi-Liang Liu, Dayiheng Liu, Dexi Liu, Fangyu Liu, Fenglin Liu, Han Liu, Haochen Liu, Haokun Liu, Haoyan Liu, Hui Liu, Jiachang Liu, Jiacheng Liu, Jian Liu, Jing Liu, Junhao Liu, Kang Liu, Lemao Liu, Ling Liu, Liyuan Liu, Ming Liu, Mingtong Liu, Nelson Liu, Peng Liu, Qian Liu, Qianchu Liu, Shujie Liu, Siyang Liu, Tianyu Liu, Wei Liu, Weijie Liu, Xianggen Liu, Xiao Liu, Xiao Liu, Xiaodong Liu, Xien Liu, Xin Liu, Xuebo Liu, Xueqing Liu, Yang Janet Liu, Yijin Liu, Yong Liu, Yongfei Liu, Yuang Liu, Zemin Liu, Zeming Liu, Zhenghao Liu, Zhengyuan Liu, Zhiyuan Liu, Zhiyue Liu, Zhun Liu, Zihan Liu, Zoey Liu, Kyle Lo, Sharid Loáiciga, Lajanugen Logeswaran, Damien Lolive, Guodong Long, Quanyu Long, Yunfei Long, Shayne Longpre, José David Lopes, Adrian Lopez monroy, Jaime Lorenzo-Trueba, Natalia Loukachevitch, Di Lu, Kaiji Lu, Wei Lu, Xing Lu, Yaojie Lu, Yichao Lu, Yu Lu, Yujie Lu, Li Lucy, Stephanie Lukin, Hongyin Luo, Huaishao Luo, Ling Luo, Ping Luo, Tianyi Luo, Kelvin Luu, Qi Lv, Shangwen Lv, Xin Lv, Qing Lyu, Meryem M'hamdi, Mathias Müller, Thomas Müller, Carlos-Francisco Méndez-Cruz, Ji Ma, Kaixin Ma, Mingyu Ma, Qianli Ma, Tingting Ma, Weicheng Ma, Xiaomeng Ma, Xinyin Ma, Xuezhe Ma, Xutai Ma, Aman Madaan, Mounica Maddela, Pranava Madhyastha, Andrea Madotto, Brielen Madureira, Manuel Mager, Saad Mahamood, Suchismit Mahapatra, Debanjan Mahata, Rahmad Mahendra, Ayush Maheshwari, Gaurav Maheshwari, Kyle Mahowald, Wolfgang Maier, Bodhisattwa Prasad Majumder, Rathnakara Malatesha, Chaitanya Malaviya, Andreas Maletti, Ankur Mali, Eric Malmi, Christopher Malon, Valentin Malykh, Saab Mansour, Ramesh Manuvinakurike, Emaad Manzoor, Wenji Mao, Xian-Ling Mao, Xin Mao, Yuning Mao, Yuren Mao, Diego Marcheggiani, Daniel Marcu, Piotr Mardziel, Benjamin Marie, Zita Marinho, Antonis Maronikolakis, Edison Marrese Taylor, Lara Martin, Marianna Martindale, Pedro Henrique Martins, Eva Martínez Garcia, Sameen Maruf, Claudia Marzi, Aleksandre Maskharashvili, Sarah Masud, Lambert Mathias, Sandeep Mathias, Puneet Mathur, David Matos, Sérgio Matos, Yevgen Matusevych, Evgeny Matusov, Nickil Maveli, Jonathan May, Stephen Mayhew, Joshua Maynez, John McCrae, Kate McCurdy, Matthew McDermott, Denis McInerney, Alexander Mehler, Shikib Mehri, Maitrey Mehta, Nikhil Mehta, Hongyuan Mei, Clara Meister, Dheeraj Mekala, Julia Mendelsohn, Telmo Menezes, Fandong Meng, Helen Meng, Rui Meng, Tao Meng, Yu Meng, Yuanliang Meng, Zhao Meng, Rakesh Menon, Wolfgang Menzel, Paola Merlo, William Merrill, Donald Metzler, Fei Mi, Haitao Mi, Qingliang Miao, Yisong Miao, Julian Michael, George Michalopoulos, Paul Michel, Lesly Miculicich, Sabrina Mielke, Zulfat Miftahutdinov, Todor Mihaylov, Tsvetomila Mihaylova, Victor Mijangos, Elena Mikhalkova, Simon Mille, Tristan Miller, Emiel Miltenburg, Eleni Miltsakaki, David Mimno, Do June Min, Sewon Min, Pasquale Minervini, Xu Mingzhou, Hideya Mino, Sabino Miranda, Paramita Mirza, Abhijit Mishra, Swaroop Mishra, Amita Misra, Teruko Mitamura, Arpit Mittal, Yusuke Miyao, Takashi Miyazaki, Daniela Moctezuma, Ashutosh Modi, Aditya Mogadala, Mahmoud Mohammadi, Alireza Mohammadshahi, Hosein Mohebbi, Diego Molla, Nicholas Monath, Ishani Mondal, Syrielle Montariol, Manuel Montes, Seungwhan Moon, Ray Mooney, Mehrad Moradshahi, Vlad Morariu, Mohamed Morchid, Jose Moreno, Mathieu Morey, Junichiro Mori, Gaku Morio, Makoto Morishita, John Morris, David Mortensen, Marius Mosbach, Aida Mostafazadeh Davani, Xiangyang Mou, Frank Mtumbuka, Hamdy Mubarak, Aaron Mueller, Shamsuddeen Muhammad, Animesh Mukherjee, Matthew Mulholland, Deepak Muralidharan, Masayasu Muraoka, Tomáš Musil, Agnieszka Mykowiecka, Sheshera Mysore, Claire Nédellec, Aurélie Névéol, Seung-Hoon Na, Masaaki Nagata, Ajay Nagesh, Suraj Nair, Saeed Najafi, Tetsuji Nakagawa, Nikita Nangia, Diane Napolitano, Jason Naradowsky, Kanika Narang, Karthik Narasimhan, Tahira Naseem, Alexis Nasr, Vivi Nastase, Anandhavelu Natarajan, Matteo Negri, Isar Nejadgholi, Preksha Nema, Graham Neubig, Günter Neumann, Mariana Neves, Denis Newman-Griffis, Jun Ping Ng, Huy Nguyen, Huyen Nguyen, Kiet Nguyen, Minh Nguyen, Minh Van Nguyen, Minh-Tien Nguyen, Thanh-Tung Nguyen, Thien Nguyen, Truc-Vien Nguyen, Hoang-Quoc Nguyen-Son, Ansong Ni, Jian Ni, Jianmo Ni, Pin Ni, Garrett Nicolai, Massimo Nicosia, Feng Nie, Shaoliang Nie, Yixin Nie, Andreas Niekler, Joel Niklaus, Giannis Nikolentzos, Vassilina Nikoulina, Lasguido Nio, Kosuke Nishida, Noriki Nishida, Masaaki Nishino, Sergiu Nisioi, Tong Niu, Cicero Nogueira dos Santos, Hiroshi Noji, Tadashi Nomoto, Farhad Nooralahzadeh, Damien Nouvel, Michal Novák, Jekaterina Novikova, Pierre Nugues, Alexander O'Connor, Tim Oates, Cennet Oguz, Byung-Doh Oh, Atul Ojha, Manabu Okumura, Eda Okur, Hugo Oliveira, Ali Omrani, Arturo Oncevay-Marcos, Ethel Ong, Yasumasa Onoe, Juri Opitz, Shereen Oraby, Aitor Ormazabal, John Ortega, Yohei Oseki, Malte Ostendorff, Myle Ott, Zebin Ou, Zhijian Ou, Zijing Ou, Nedjma Ousidhoum, Andrew Owens, Deepak P, Juan Antonio Pérez-Ortiz, Maria Pacheco, Inkit Padhi, Vishakh Padmakumar, Aline Paes, Vardaan Pahuja, Kuntal Pal, Santanu Pal, Chester Palen-Michel, Alonso Palomino, Liangming Pan, Liang Pang, Yuanzhe Pang, Alexandros Papangelis, Raghavendra Pappagari, Nikolaos Pappas, Emerson Paraiso, Georgios Paraskevopoulos, Letitia Parcalabescu, Natalie Parde, Zarana Parekh, Cecile Paris, ChaeHun Park, Chanjun Park, Eunjeong Park, Gilchan Park, Hyunji Park, Jungsoo Park, Kunwoo Park, Youngja Park, Alicia Parrish, Md Rizwan Parvez, Alexandre Passos, Ramakanth Pasunuru, Arkil Patel, Raj Patel, Sangameshwar Patil, Barun Patra, Braja Patra, Jasabanta Patro, Manasi Patwardhan, Siddharth Patwardhan, Shachi Paul, Ellie Pavlick, John Pavlopoulos, Pavel Pecina, Stephan Peitz, Viktor Pekar, Baolin Peng, Haoruo Peng, Siyao Peng, Xi Peng, Xutan Peng, Yifan Peng, Lis Pereira, Martin Pereira, Olatz Perez-de-Vinaspre, Gabriele Pergola, Jan-Thorsten Peter, Ben Peters, Matthew Peters, Pavel Petrushkov, Jonas Pfeiffer, Minh Pham, Quan Pham, Van-Thuy Phi, Maciej Piasecki, Massimo Piccardi, Karl Pichotta, Matúš Pikuliak, Tiago Pimentel, Aidan Pine, Juan Pino, Yuval Pinter, Tommi Pirinen, Rajesh Piryani, Nikiforos Pittaras, Lidia Pivovarova, Benjamin Piwowarski, Brian Plüss, Peter Plantinga, Bryan Plummer, Lahari Poddar, Massimo Poesio, Heather Pon-Barry, Martin Popel, Octavian Popescu, Andrei Popescu-Belis, Ian Porada, Matt Post, Martin Potthast, Christopher Potts, Amir Pouran Ben Veyseh, Aniket Pramanick, Jakob Prange, Animesh Prasad, Archiki Prasad, Adithya Pratapa, Judita Preiss, Luigi Procopio, Victor Prokhorov, Prokopis Prokopidis, Haritz Puerto, Jay Pujara, Rajkumar Pujari, Hemant Purohit, Matthew Purver, Ehsan Qasemi, Fan-chao Qi, Jianzhong Qi, Peng Qi, Tao Qi, Dong Qian, Jing Qian, Kun Qian, Yujie Qian, Bowen Qin, Lianhui Qin, Libo Qin, Wenda Qin, Yujia Qin, Long Qiu, Xipeng Qiu, Yunqi Qiu, Zimeng Qiu, Chen Qu, Xiaoye Qu, Andreas Rücklé, Ella Rabinovich, Gorjan Radevski, Alessandro Raganato, Ash Rahimi, Hossein Rajaby Faghihi, Sara Rajaee, Dheeraj Rajagopal, Sanguthevar Rajasekaran, Pavithra Rajendran, Geetanjali Rakshit, Dhananjay Ram, Ori Ram, Owen Rambow, Abhinav Ramesh Kashyap, Alan Ramponi, Gabriela Ramírez de la Rosa, Tharindu Ranasinghe, Sudarshan Rangarajan, Priya Rani, Peter Rankel, Dongning Rao, Jinfeng Rao, Sudha Rao, Ahmad Rashid, Hannah Rashkin, Vikas Raunak, Shauli Ravfogel, Vinit Ravishankar, Anirudh Ravula, Baishakhi Ray, Soumya Ray, Julia Rayz, Traian Rebedea, Sravana Reddy, Brian Reese, Marek Rei, Nils Reimers, Navid Rekabsaz, Da Ren, Feiliang Ren, Feiliang Ren, Pengjie Ren, Ruiyang Ren, Shuhuai Ren, Shuo Ren, Xiang Ren, Xingzhang Ren, Zhaochun Ren, Adithya Renduchintala, Robert Reynolds, Mehdi Rezagholizadeh, Rezvaneh Rezapour, Saed Rezayi, Yoo Rhee Oh, Maksim Riabinin, Leonardo Ribeiro, Marco Tulio Ribeiro, Caitlin Richter, Sebastian Riedel, Mark Riedl, Stefan Riezler, German Rigau, Matīss Rikters, Darcey Riley, Annette Rios, Anthony Rios, Miguel Rios Gaona, Elijah Rippeth, Brian Roark, Kirk Roberts, Christophe Rodrigues, Paul Rodrigues, Juan Rodriguez, Melissa Roemmele, Paul Roit, Lina Rojas-Barahona, Stephen Roller, Alexey Romanov, Salvatore Romeo, Subendhu Rongali, Weng Rongxiang, Michael Roth, Bryan Routledge, Subhro Roy, Sharod Roy Choudhury, Jos Rozen, Alla Rozovskaya, Dongyu Ru, Raphael Rubino, Sebastian Ruder, Koustav Rudra, Frank Rudzicz, Federico Ruggeri, Josef Ruppenhofer, Thomas Ruprecht, Alexander Rush, Phillip Rust, Piotr Rybak, Maria Ryskina, Masoud Sabet, Ashish Sabharwal, Mrinmaya Sachan, Fatiha Sadat, Arka Sadhu, Niloofar Safi Samghabadi, Benoît Sagot, Monjoy Saha, Swarnadeep Saha, Tulika Saha, Saurav Sahay, Gaurav Sahu, Hassan Sajjad, Keisuke Sakaguchi, Hiroki Sakaji, Sakriani Sakti, Jonne Saleva, Avneesh Saluja, Bidisha Samanta, Tanja Samardzic, Younes Samih, Shailaja Keyur Sampat, David Samuel, Abhilasha Sancheti, Vicente Sanchez Carmona, Erik Sang, Chinnadhurai Sankar, Sashank Santhanam, Marina Santini, Bishal Santra, Maarten Sap, Naomi Saphra, Maya Sappelli, Sujoy Sarkar, Sheikh Muhammad Sarwar, Giorgio Satta, Danielle Saunders, Beatrice Savoldi, Ramit Sawhney, Apoorv Saxena, Bianca Scarlini, Shigehiko Schamoni, Yves Scherrer, Timo Schick, Frank Schilder, Viktor Schlegel, Helmut Schmid, Tyler Schnoebelen, Steven Schockaert, Claudia Schulz, Elliot Schumacher, Anne-Kathrin Schumann, Tal Schuster, Robert Schwarzenberg, Stefan Schweter, Djamé Seddah, João Sedoc, Amit Seker, Thibault Sellam, David Semedo, Nasredine Semmar, Indira Sen, Lütfi Kerem Senel, Minjoon Seo, Yeon Seonwoo, Prashant Serai, Sofia Serrano, Christophe Servan, Karin Sevegnani, Silvia Severini, Lei Sha, Izhak Shafran, Samira Shaikh, Cory Shain, Igor Shalyminov, Zain Shamsi, Chao Shang, Guokan Shang, Mingyue Shang, Chenze Shao, Nan Shao, Yutong Shao, Zhihong Shao, Ori Shapira, Matthew Shardlow, Ehsan Shareghi, Arpit Sharma, Ashish Sharma, Vasu Sharma, Serge Sharoff, Rebecca Sharp, Hassan Shavarani, Peter Shaw, Shahin Shayandeh, Ravi Shekhar, Artem Shelmanov, Aili Shen, Hua Shen, Jiaming Shen, Lei Shen, Qinlan Shen, Sheng Shen, Shiqi Shen, Tao Shen, Xiaoyu Shen, Yikang Shen, Yongliang Shen, Emily Sheng, Qiang Sheng, Thomas Sherborne, Chen Shi, Freda Shi, Jiatong Shi, Jiaxin Shi, Ning Shi, Peng Shi, Shuming Shi, Weijia Shi, Weiyan Shi, Xing Shi, Yangyang Shi, Chihiro Shibata, Takashi Shibuya, Anastasia Shimorina, Jamin Shin, Kazutoshi Shinoda, Keiji Shinzato, Yow-Ting Shiue, Segev Shlomov, Eyal Shnarch, Abu Awal Md Shoeb, Ritvik Shrivastava, Kai Shu, Lei Shu, Raphael Shu, Kurt Shuster, Vered Shwartz, Chenglei Si, Mei Si, A.b. Siddique, Carina Silberer, Miikka Silfverberg, Fabrizio Silvestri, Kathleen Siminyu, Dan Simonson, Edwin Simpson, Arabella Sinclair, Chandan Singh, Kuldeep Singh, Mayank Singh, Karan Singla, Koustuv Sinha, Kairit Sirts, Amy Siu, Milena Slavcheva, Noam Slonim, David Smith, Marco Sobrevilla Cabezudo, Hyun-Je Song, Kai Song, Kaiqiang Song, Linfeng Song, Linqi Song, Mingyang Song, Siqi Song, Wei Song, Xingyi Song, Sandeep Soni, Rishi Sonthalia, Claudia Soria, Alexey Sorokin, Sajad Sotudeh, José Souza, Alexander Spangher, Matthias Sperber, Daniel Spokoyny, Balaji Vasan Srinivasan, Tejas Srinivasan, Edward Stabler, Felix Stahlberg, Ieva Staliunaite, Marija Stanojevic, Katherine Stasaski, Manfred Stede, Mark Steedman, Julius Steen, Shane Steinert-Threlkeld, Elias Stengel-Eskin, Samuel Stevens, Symon Stevens-Guille, Mark Stevenson, Ian Stewart, Matthew Stone, Kevin Stowe, Kristina Striegnitz, Nikolaos Stylianou, Dan Su, Sheng Su, Weifeng Su, Yixuan Su, Yu Su, Nishant Subramani, Katsuhito Sudoh, Hiroaki Sugiyama, Kazunari Sugiyama, Alessandro Suglia, Yoshihiko Suhara, Dianbo Sui, Zhifang Sui, Changzhi Sun, Chengjie Sun, Haipeng Sun, Huan Sun, Jian Sun, Kai Sun, Kai Sun, Ming Sun, Simeng Sun, Siqi Sun, Tianxiang Sun, Xiaobing Sun, Yibo Sun, Yu Sun, Zequn Sun, Zhiqing Sun, Dhanasekar Sundararaman, Yun-Hsuan Sung, Hanna Suominen, Anshuman Suri, Simon Suster, Mirac Suzgun, Sandesh Swamy, Reid Swanson, Swabha Swayamdipta, Benjamin Sznajder, Stan Szpakowicz, Gözde Şahin, Ryuki Tachibana, Oyvind Tafjord, Chang-Yu Tai, Hagai Taitelbaum, Hiroya Takamura, Sho Takase, Ece Takmaz, Derek Tam, Ronen Tamari, Fabio Tamburini, Akihiro Tamura, Chuanqi Tan, Fei Tan, Liling Tan, Samson Tan, Xu Tan, Kumiko Tanaka-Ishii, Hao Tang, Jiliang Tang, Liyan Tang, Raphael Tang, Shuai Tang, Siliang Tang, Yi-Kun Tang, Zhiwen Tang, Chongyang Tao, Shiva Taslimipoor, Yuka Tateisi, Marta Tatu, Hillel Taub-Tabib, Stephen Taylor, Selma Tekir, Serra Tekiroglu, Irina Temnikova, Zhiyang Teng, Ian Tenney, Maartje Ter Hoeve, Hiroki Teranishi, Hrishikesh Terdalkar, Silvia Terragni, Alberto Testoni, Simone Teufel, Nithum Thain, Ashish Thapliyal, Mokanarangan Thayaparan, Stephen Thomas, Jesse Thomason, Sam Thomson, Camilo Thorne, James Thorne, Chunwei Tian, Junfeng Tian, Ran Tian, Yingtao Tian, Zhiliang Tian, Jörg Tiedemann, Tiago Timponi Torrent, Evgeniia Tokarchuk, Ryoko Tokuhisa, Nadi Tomeh, Nicholas Tomlin, Sara Tonelli, Mariya Toneva, Kentaro Torisawa, Juan-Manuel Torres-Moreno, Samia Touileb, Julien Tourille, Khanh Tran, Thy Tran, Marcos Treviso, Enrica Troiano, Adam Tsakalidis, Bo-Hsiang Tseng, Yuen-Hsien Tseng, Masaaki Tsuchida, Yoshimasa Tsuruoka, Mei Tu, Zhaopeng Tu, Nicolas Turenne, Martin Tutek, Francis Tyers, Ana Uban, Takuma Udagawa, Dennis Ulmer, Shyam Upadhyay, Tanguy Urvoy, Raúl Vázquez, Sowmya Vajjala, Josef Valvoda, Jannis Vamvas, Tim Van de Cruys, Vincent Vandeghinste, Lucy Vanderwende, David Vandyke, Natalia Vanetik, Daniel Varab, Siddharth Vashishtha, Julien Velcin, Sriram Venkatapathy, Suzan Verberne, Gaurav Verma, Rakesh Verma, Giorgos Vernikos, Prashanth Vijayaraghavan, David Vilar, Jesús Vilares, Martin Villalba, Veronika Vincze, Krishnapriya Vishnubhotla, Rob Voigt, Elena Voita, Pooja Voladoddi, Soroush Vosoughi, Son Vu, Thang Vu, Thuy-Trang Vu, Tu Vu, Yogarshi Vyas, Carel van Niekerk, Rik van Noord, Jan-Willem van de Meent, Henning Wachsmuth, Takashi Wada, David Wadden, Sabine Walde, Andreas Waldis, Mengting Wan, Mingyu Wan, Yao Wan, Yu Wan, Alex Wang, B in Wang, Bailin Wang, Baoxin Wang, Baoxun Wang, Bingqing Wang, Boxin Wang, Chenguang Wang, Chengyu Wang, Cunxiang Wang, Daling Wang, Fei Wang, Guangrun Wang, Guoxin Wang, Guoyin Wang, Hai Wang, Han Wang, Hanrui Wang, Hao Wang, Haoyu Wang, Haoyu Wang, Hong Wang, Hongfei Wang, Hua Wang, Jin Wang, Jin Wang, Jingang Wang, Jue Wang, Lei Wang, Liang Wang, Lidan Wang, Lingzhi Wang, Mengxiang Wang, Ping Wang, Qiang Wang, Qingyun Wang, Quan Wang, Rui Wang, Rui Wang, Runze Wang, Shaonan Wang, Shi Wang, Shuo Wang, Shuohang Wang, Sinong Wang, Siyuan Wang, Tong Wang, Tong Wang, Wei Wang, Weiyue Wang, Wenhui Wang, Wenya Wang, Xiaojie Wang, Xiaolin Wang, Xiaorui Wang, Xin Wang, Xing Wang, Xintong Wang, Xinyiw Wang, Xuezhi Wang, Yan Wang, Yaqing Wang, Yaqing Wang, Ye Wang, Yequan Wang, Yifei Wang, Yijue Wang, Yingyao Wang, Yiran Wang, Yizhong Wang, Yue Wang, Yue Wang, Yufei Wang, Yujing Wang, Zekun Wang, Zhefeng Wang, Zhen Wang, Zhenyi Wang, Zhichun Wang, Zhongqing Wang, Zijian Wang, Zirui Wang, Wei Wang., Nigel Ward, Alex Warstadt, Christian Wartena, Koki Washio, Ingmar Weber, Leon Weber, Albert Webson, Jason Wei, Junqiu Wei, Penghui Wei, Wei Wei, Xiangpeng Wei, Shira Wein, Nathaniel Weir, Ralph Weischedel, Charles Welch, Orion Weller, Haoyang Wen, Lijie Wen, Peter West, Taesun Whang, Michael Wiegand, Sarah Wiegreffe, Adam Wiemerslage, Derry Wijaya, Gijs Wijnholds, Jake Williams, Jennifer Williams, Steven Wilson, Genta Winata, Shuly Wintner, Sam Wiseman, Guillaume Wisniewski, Tomer Wolfson, Derek Wong, Tak-Lam Wong, Bobby Wu, Bowen Wu, Chien-Sheng Wu, Chuhan Wu, Fangzhao Wu, Felix Wu, Jian Wu, Junshuang Wu, Lianwei Wu, Qianhui Wu, Qingyang Wu, Shuangzhi Wu, Sixing Wu, Stephen Wu, Wei Wu, Wenhao Wu, Xianchao Wu, Xiaobao Wu, Xixin Wu, Yanan Wu, Yiquan Wu, Yu Wu, Yuanbin Wu, Yunfang Wu, Yuting Wu, Yuxiang Wu, Zeqiu Wu, Zhanghao Wu, Zhengxuan Wu, Zhiyong Wu, Zhonghai Wu, Ai Xi, Congying Xia, Jingbo Xia, Patrick Xia, Qingrong Xia, Rui Xia, Yandi Xia, Yikun Xian, Jiannan Xiang, Rong Xiang, Bo Xiao, Chaojun Xiao, Chunyang Xiao, Jinghui Xiao, Lin Xiao, Liqiang Xiao, Min Xiao, Wen Xiao, Yanghua Xiao, Jun Xie, Qianqian Xie, Ruobing Xie, Tianbao

Xie, Yuqiang Xie, Linzi Xing, Wenhan Xiong, Benfeng Xu, Boyan Xu, Can Xu, Canwen Xu, Chen Xu, Frank Xu, Jia Xu, Jiacheng Xu, Jing Xu, Jingjing Xu, Jitao Xu, Jun Xu, Kun Xu, Kun Xu, Lin Xu, Liyan Xu, Peng Xu, Peng Xu, Qiantong Xu, Qiongkai Xu, Ruifeng Xu, Runxin Xu, Ruochen Xu, Shusheng Xu, Wang Xu, Weijia Xu, Weiran Xu, Wenduan Xu, Xinnuo Xu, Yan Xu, Yige Xu, Yumo Xu, Zenglin Xu, Zhen Xu, Zhiyang Xu, Mohit Yadav, Prateek Yadav, Yadollah Yaghoobzadeh, Ivan Yamshchikov, An Yan, Jun Yan, Rui Yan, Xu Yan, Yu Yan, Yuanmeng Yan, Baosong Yang, Changbing Yang, Fan Yang, Haiqin Yang, Jingfeng Yang, Jun Yang, Junjie Yang, Liner Yang, Linyi Yang, Min Yang, Ruosong Yang, Sen Yang, Songlin Yang, Tsung-Yen Yang, Wei Yang, Wenmian Yang, Xiaoyu Yang, Yilin Yang, Yinfei Yang, Yujiu Yang, Zhao Yang, Zhen Yang, Zhixian Yang, Ziyi Yang, Huaxiu Yao, Jianmin Yao, Liang Yao, Liang Yao, Shunyu Yao, Wenlin Yao, Ziyu Yao, Mark Yatskar, Deming Ye, Qinyuan Ye, Wei Ye, Xi Ye, Reyyan Yeniterzi, Jinyoung Yeo, Xiaoyuan Yi, Olcay Yildiz, Wen-wai Yim, Seid Yimam, Da Yin, Fan Yin, Jian Yin, Kayo Yin, Pengcheng Yin, Qingyu Yin, Wenpeng Yin, Yichun Yin, Michael Yoder, Sho Yokoi, Wang Yong, Jin Yong Yoo, Kang Min Yoo, Hiyori Yoshikawa, Weiqiu You, Steve Young, Bei Yu, Bowen Yu, Changlong Yu, Chen Yu, Dian Yu, Dian Yu, Dong Yu, Donghan Yu, Heng Yu, Juntao Yu, Liang-Chih Yu, Seunghak Yu, Sicheng Yu, Tao Yu, Wenhao Yu, Xiaodong Yu, Yue Yu, Zhang Yu, Zhiwei Yu, Caixia Yuan, David Yuan, Jianhua Yuan, Nicholas Jing Yuan, Weizhe Yuan, Xiang Yue, Hyokun Yun, Matrix Z, Wajdi Zaghouani, Hamada Zahera, Hamed Zamani, Sina Zarrieß, Rabih Zbib, Albin Zehe, Rowan Zellers, Yury Zemlyanskiy, Daojian Zeng, Jiali Zeng, Jichuan Zeng, Qi Zeng, Qingkai Zeng, Shuang Zeng, Weixin Zeng, Xingshan Zeng, Zhiyuan Zeng, Deniz Zeyrek, Hanwen Zha, Fangzhou Zhai, Haolan Zhan, Bowen Zhang, Chao Zhang, Chen Zhang, Chen Zhang, Chiyu Zhang, Chuheng Zhang, Danqing Zhang, Dejiao Zhang, Denghui Zhang, Dong Zhang, Dongdong Zhang, Guanhua Zhang, Haibo Zhang, Haisong Zhang, Hanlei Zhang, Hanlin Zhang, Hao Zhang, Haoyu Zhang, Hongming Zhang, Huijun Zhang, Jiajun Zhang, Jianguo Zhang, Jiayao Zhang, Jingqing Zhang, Ke Zhang, Kun Zhang, Lei Zhang, Lei Zhang, Licheng Zhang, Liwen Zhang, Longyin Zhang, Meishan Zhang, Meng Zhang, Michael Zhang, Mike Zhang, Min Zhang, Ningyu Zhang, Peng Zhang, Qi Zhang, Qiang Zhang, Richong Zhang, Ruixiang Zhang, Ruiyi Zhang, Sheng Zhang, Shiyue Zhang, Shujian Zhang, Shuo Zhang, Songyang Zhang, Tianlin Zhang, Tianyun Zhang, Tong Zhang, Tongtao Zhang, Wei Emma Zhang, Wen Zhang, Xiang Zhang, Xiao Zhang, Xiaotong Zhang, Xiaoying Zhang, Xinliang Frederick Zhang, Xinsong Zhang, Xinyuan Zhang, Xuan Zhang, Xuanyu Zhang, Xuchao Zhang, Yan Zhang, Yan Zhang, Yao Zhang, Yichi Zhang, Yu Zhang, Yuan Zhang, Yuanzhe Zhang, Yuhui Zhang, Yunyan Zhang, Yunyi Zhang, Yuqi Zhang, Zeyu Zhang, Zheng Zhang, Zhengyan Zhang, Zhirui Zhang, Zhisong Zhang, Zhiwei Zhang, Zhong Zhang, Zhuosheng Zhang, Ziqi Zhang, Chao Zhao, Chen Zhao, Dongyan Zhao, Guangxiang Zhao, Jie Zhao, Jieyu Zhao, Kai Zhao, Mengjie Zhao, Sanqiang Zhao, Tiancheng Zhao, Tianyu Zhao, Tuo Zhao, Yang Zhao, Yao Zhao, Yilun Zhao, Zhenjie Zhao, Bo Zheng, Changmeng Zheng, Chujie Zheng, Renjie Zheng, Xiaoqing Zheng, Yinhe Zheng, Zaixiang Zheng, Ming Zhong, Peixiang Zhong, Victor Zhong, Wanjun Zhong, Zexuan Zhong, Ben Zhou, Chunting Zhou, Deyu Zhou, Dong Zhou, Giulio Zhou, Guangyou Zhou, Jiawei Zhou, Jie Zhou, Jie Zhou, Jingbo Zhou, Junpei Zhou, Li Zhou, Long Zhou, Pei Zhou, Qingyu Zhou, Wangchunshu Zhou, Wenxuan Zhou, Xiang Zhou, Xiangyang Zhou, Xuhui Zhou, Yaqian Zhou, Yi Zhou, Yichu Zhou, Yilun Zhou, Yucheng Zhou, Zhengyu Zhou, Zhihan Zhou, Zhixuan Zhou, Conghui Zhu, Hao Zhu, Jian Zhu, Junnan Zhu, Kenny Zhu, Ligeng Zhu, Lixing Zhu, Muhua Zhu, Qingfu Zhu, Qinglin Zhu, Su Zhu, Wei Zhu, Xiaofeng Zhu, Yilun Zhu, Yong Zhu, Yutao Zhu, Zining Zhu, Fuzhen Zhuang, Yimeng Zhuang, Caleb Ziems, Roger Zimmermann, Ayah Zirikly, Yftah Ziser, Shi Zong, Bowei Zou, Amal Zouaq, Arkaitz Zubiaga, Pierre Zweigenbaum

Emergency Reviewers

Sweta Agrawal, Hend Al-Khalifa, Tariq Alhindi, Xiang Ao, Ignacio Arroyo-Fernández, Eleftherios Avramidis, Fan Bai, Niranjan Balasubramanian, Daniel Bauer, Adrian Benton, Yonatan Bisk, Logan Born, Piyush Borole, Nadjet Bouayad-Agha, Florian Boudin, Stephanie Brandl, Chris

Brew, Eleftheria Briakou, Joan Byamugisha, Rémi Cardon, Khyathi Chandu, Guanhua Chen, Hanjie Chen, Hong Chen, Lei Chen, Qian Chen, Wang Chen, Zifeng Cheng, Ethan Chi, Yew Ken Chia, Prafulla Kumar Choubey, Brian Davis, Xiang Deng, Aleksandr Drozd, Dylan Ebert, Micha Elsner, Martin Fajčík, Amir Feder, Xiaocheng Feng, Javier Ferrando, Francis Ferraro, Simone Filice, Lea Frermann, Xingyu Fu, Jana Götze, Yang Gao, Deepanway Ghosal, Sucheta Ghosh, Daniel Gildea, Goran Glavaš, Samuel Gonzalez-Lopez, Yulia Grishina, Jing Gu, Jiaqi Guo, Yinpeng Guo, David Hall, Rujun Han, Taiqi He, Ryuichiro Higashinaka, Mark Hopkins, Benjamin Hsu, Wei Hu, Danqing Huang, Chia-Chien Hung, Dieuwke Hupkes, Mert Inan, Sujay Kumar Jauhar, Tommi Jauhiainen, Lifeng Jin, Dhanya Jothimani, Mladen Karan, Börje Karlsson, Yafeng Ken, Ekaterina Kochmar, Yuta Koreeda, Udo Kruschwitz, Yash Kumar Lal, Wai Lam, Wuwei Lan, Md Tahmid Rahman Laskar, Brian Lester, Ashley Lewis, Mike Lewis, Dingcheng Li, Dongfang Li, Jingjing Li, Liangyou Li, Manling Li, Yuliang Li, Lizi Liao, Robert Litschko, Lemao Liu, Yang Janet Liu, Yuang Liu, Zhenghao Liu, Zihan Liu, Adrian Pastor Lopez Monroy, Yu Lu, Yujie Lu, Xin Lv, Qing Lyu, Carlos-Francisco Méndez-Cruz, Manuel Mager, Suchismit Mahapatra, Rathnakara Malatesha, Eva Martínez Garcia, Shikib Mehri, Nikhil Mehta, Hongyuan Mei, Victor Mijangos, Sewon Min, Xu Mingzhou, Sabino Miranda, Teruko Mitamura, Ashutosh Modi, Aditya Mogadala, Mohamed Morchid, Lili Mou, Xiangyang Mou, Suraj Nair, Roberto Navigli, Mariana Neves, Andreas Niekler, Byung-Doh Oh, Ethel Ong, Aitor Ormazabal, Zebin Ou, Maria Pacheco, Vardaan Pahuja, Alonso Palomino, Zarana Parekh, Ellie Pavlick, Jonas Pfeiffer, Van-Thuy Phi, Lidia Pivovarova, Christopher Potts, Dong Qian, Gorjan Radevski, Abhinav Ramesh Kashyap, Dongning Rao, Sebastian Riedel, Elijah Rippeth, Brian Roark, Kirk Roberts, Sharod Roy Choudhury, Alexander Rush, Monjoy Saha, Abhilasha Sancheti, Shahin Shayandeh, Yangyang Shi, Yow-Ting Shiue, Segev Shlomov, Fabrizio Silvestri, Linfeng Song, Linqi Song, Marija Stanojevic, Samuel Stevens, Symon Stevens-Guille, Huan Sun, Jian Sun, Jian Sun, Zeerak Tala, Liyan Tang, Chunwei Tian, Khanh Tran, Tanguy Urvoy, Siyuan Wang, Ye Wang, Yue Wang, Zhen Wang, Christian Wartena, Guillaume Wisniewski, Jian Wu, Qingyang Wu, Xixin Wu, Zhengxuan Wu, Weijia Xu, Wenduan Xu, Yan Xu, Yadollah Yaghoobzadeh, Jun Yang, Linyi Yang, Jin-Ge Yao, Wenlin Yao, Wenpeng Yin, Bowen Yu, Weizhe Yuan, Guanhua Zhang, Jiayao Zhang, Qiang Zhang, Xiang Zhang, Xiaoying Zhang, Yichi Zhang, Zeyu Zhang, Zhixuan Zhou, Shi Zong

Outstanding Action Editors

Alla Rozovskaya, City University of New York
Ankur Parikh, Google
Byron Wallace, Northeastern University
Chloe Braud, CNRS
Daniel Preotiuc-Pietro, Bloomberg
Diarmuid Séaghdha, Apple
Edoardo Ponti, MILA - Quebec AI Institute
Ekaterina Kochmar, University of Bath
Elena Cabrio, Université Côte d'Azur
Emma Strubell, Carnegie Mellon University
François Yvon, LISN-CNRS / Université Paris Saclay
Ganesh Ramakrishnan, Indian Institute of Technology Bombay Indian Institute of Technology Bombay
Hyeju Jang, University of British Columbia
John Conroy, Institute for Defense Analyses Center for Computing Sciences
Kokil Jaidka, National University of Singapore
Li Dong, Microsoft Research
Longyue Wang, Tencent AI Lab
Maciej Ogrodniczuk, Institute of Computer Science Polish Academy of Sciences
Margot Mieskes, University of Applied Sciences Darmstadt

Maria Liakata, Queen Mary University London
Marianna Apidianaki, University of Pennsylvania University of Pennsylvania
Michael Zock, CNRS
Mikel Artetxe, Facebook AI Research
Miles Osborne, Epistemic.ai
Nathan Schneider, Georgetown University
Raj Dabre, National Institute of Information and Communications Technology (NICT) National Institute of Advanced Industrial Science and Technology
Sara Stymne, Uppsala University
Timothy Baldwin, The University of Melbourne
Tong Xiao, Northeastern University
Trang Tran, USC Institute for Creative Technologies University of Southern California
Xin Wang, University of California Santa Cruz
Yonatan Bisk, Carnegie Mellon University
Zhe Gan, Microsoft

Outstanding Reviewers

Aaron Mueller, Johns Hopkins University
Abram Handler, University of Colorado at Boulder
Aida Mostafazadeh Davani, University of Southern California
Albin Zehe, University of Würzburg
Amanda Stent, Colby College
Andrew Lee, University of Michigan
Aniket Pramanick, Technische Universität Darmstadt
Arabella Sinclair, University of Amsterdam
B in Wang, National University of Singapore
Benjamin Marie, National Institute of Information and Communications Technology (NICT) National Institute of Advanced Industrial Science and Technology
Brendan Kennedy, University of Southern California
Brian Roark, Google
Christopher Clark, Allen Institute for Artificial Intelligence
Christos Christodoulopoulos, Amazon
Clara Meister, Swiss Federal Institute of Technology
Dallas Card, University of Michigan - Ann Arbor
Dan Simonson, BlackBoiler Inc.
Daniel Varab, IT University of Copenhagen
Darcey Riley, University of Notre Dame
David Demeter, Northwestern University
Dayne Freitag, SRI International
Deepak P, Queen's University Belfast
Denis Newman-Griffis, University of Sheffield
Do June Min, University of Michigan - Ann Arbor
Donald Metzler, Google
Eleftheria Briakou, Department of Computer Science University of Maryland College Park
Elias Stengel-Eskin, Johns Hopkins University
Elisa Ferracane, Abridge AI
Ella Rabinovich, University of Toronto
Ellie Pavlick, Brown University
Hannah Rashkin, Google
Ian Stewart, University of Michigan
Jad Kabbara, McGill University
Jan-Thorsten Peter, AppTek

Jie Gao, University of Sheffield
Johannes Kiesel, Bauhaus-Universität Weimar
Jose Camacho-Collados, Cardiff University
Junjie Yang, Télécom ParisTech
Kartik Goyal, Toyota Technological Institute at Chicago
Ken Barker, International Business Machines
Kyle Lo, Allen Institute for Artificial Intelligence
Lara Martin, Department of Computer and Information Science School of Engineering and Applied Science
Liling Tan, Amazon
Longyin Zhang, Soochow University
Luisa Bentivogli, Fondazione Bruno Kessler
Lütfi Kerem Senel, Ludwig Maximilian University of Munich
Matthew Lease, Amazon
Melissa Roemmele, Language Weaver (RWS)
Michal Novák, Charles University Prague
Neele Falk, University of Stuttgart University of Stuttgart
Nikhil Mehta, Purdue University
Paul Michel, School of Computer Science Carnegie Mellon University
Paul Roit, Bar-Ilan University
Philip Gorinski, Huawei Noah's Ark Lab
Pierre Colombo, CentraleSupelec
Qiang Sheng, Institute of Computing Technology Chinese Academy of Sciences
Rahul Gupta, Amazon
Rebecca Knowles, National Research Council Canada
Reno Kriz, Johns Hopkins University
Saadia Gabriel, University of Washington
Serina Chang, Stanford University
Siddharth Dalmia, School of Computer Science Carnegie Mellon University
Simone Conia, Sapienza University of Rome
Sofia Serrano, University of Washington
Stan Szpakowicz, University of Ottawa
Sudarshan Rangarajan, International Business Machines
Sudha Rao, Microsoft
Tanya Goyal, University of Texas Austin
Tobias Falke, Amazon
Tomer Wolfson, Tel Aviv University
Tommi Jauhiainen, University of Turku
Trevor Cohen, University of Washington
Tristan Miller, Austrian Research Institute for Artificial Intelligence
Verna Dankers, University of Edinburgh
Xinliang Frederick Zhang, University of Michigan
Yang Deng, The Chinese University of Hong Kong
Yow-Ting Shiue, Department of Computer Science University of Maryland College Park
Yu Cao, University of Sydney
Yuntian Deng, Harvard University
Yuta Koreeda, Hitachi America Ltd.
Yves Scherrer, University of Helsinki

2

Anti-Harassment Policy

NAACL 2022 adheres to the ACL Anti-Harassment Policy. Any participant who experiences harassment or hostile behaviour may contact any current member of the ACL Professional Conduct Committee or Priscilla Rasmussen, who is usually available at the registration desk of the conference. Please be assured that if you approach us, your concerns will be kept in strict confidence, and we will consult with you on any actions taken.

The open exchange of ideas, the freedom of thought and expression, and respectful scientific debate are central to the aims and goals of a ACL conference. 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 participants at our events and in our programs.

Harassment and hostile behavior are unwelcome at any ACL conference. This includes: speech or behavior (including in public presentations and on-line discourse) that intimidates, creates discomfort, or interferes with a person's participation or opportunity for participation in the conference. We aim for ACL conferences 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, 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 ACL board 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`

3

Meal Info

A light Breakfast will be provided daily and can be found on levels 3 & 5 pre-function. During breaks, tea, coffee, pastries and snacks will be provided early morning, mid-morning and mid-afternoon and will be found in the Regency Ballroom on level 7 (on Sunday the location will be level 3 pre-function). Lunch is not provided, but there are plenty of cafes, restaurants and shops near by within walking distance as well as the hotel has a grab and go market open until 14:00. Dinner will only be provided during the Welcome Reception on Sunday evening (starting at 18:30) located on Level 5 pre-function and Tuesday evening at the social event at MoPop.

4 Social Events

We have planned the following social events for NAACL 2022. Please follow ACL's code of conduct and COVID policy when you are attending these events.

Social Event - **Tuesday, July 12th, 2022**

Venue: **MoPop Museum**
Time: **19.00 - 22.00**

We are having a social event at Museum of Pop Culture (MoPop) is a nonprofit museum in Seattle, Washington, dedicated to contemporary popular culture, MoPOP is home to the world's most immersive pop culture experiences, showcasing iconic moments in TV, music, queer culture, science fiction, and much more with collections from shoegaze and dreampop.

One Entrance ticket will be included with each main conference registration so make sure to place it in your wallet, your ticket is your entrance in, no ticket no entrance. We may have a few extra tickets for sale for accompanying persons so, if you are interested, stop by the registration desk when there are no lines.

Museum of Pop Culture is located on the Seattle Center campus, 325 5th Ave N, Seattle, WA 98105

Getting there: Your in-person Registration includes a round trip Monorail ticket that you will receive at the registration. For context, Seattle Center Monorail provides a fast, direct connection between downtown Seattle and Seattle Center traveling 25 feet above street traffic. The train departs every 10 minutes. You'll board at the Westlake Station (located downtown at 5th and Pine) and take it to Seattle Center. The museum is adjacent to the station.

5 Keynotes

Shaping Technology with Moral Imagination: Leveraging the Machinery of Value Sensitive Design

Batya Friedman
Information School, University of Washington

Monday, July 11, 2022 - Room: **Columbia A/C/D & 302 Beckler** - Time: **9:15-10:15**

Abstract: Tools and technologies are fundamental to the human condition. They do no less than create and structure the conditions in which we live, express ourselves, enact society, and experience what it means to be human. They are also the result of our moral and technical imaginations. Yet, with our limited view, it is not at all obvious how to design and engineer tools and technology so that they are more likely to support the actions, relationships, institutions, and experiences that human beings care deeply about – a life and society of human flourishing.

Value Sensitive Design (VSD) was developed as an approach to address this challenge from within technical design processes. Drawing on over three decades of work, in this plenary talk I will provide an introduction to value sensitive design foregrounding human values in the technical design process. My remarks will present some of value sensitive design's core theoretical constructs. Along the way, I'll provide some examples of applying value sensitive design to robots for healthcare and to bias in computing systems as well as demonstrate one toolkit—The Envisioning Cards—in the context of a design activity.

As time permits, I will turn to a discussion of structure, scale and time: we act within existing structure in the now, from which futures unfold across time and scale. I will unpack these observations and their implications for artificial intelligence and machine learning technologies. Thinking longer-term and systemically, I will bring forward a range of potential challenges and offer some constructive ways forward. My comments will engage individual lives, society writ large, what it means to be human, the planet and beyond.

Please have scratch paper and a pencil handy for the design activity.

Bio: Batya Friedman is a Professor in the Information School and holds adjunct appointments in the Paul G. Allen School of Computer Science & Engineering, the School of Law, and the Department of Human Centered Design and Engineering at the University of Washington where she co-founded the Value Sensitive Design Lab and the UW Tech Policy Lab. Dr. Friedman pioneered value sensitive design (VSD), an established approach to account for human values in the design of technical systems. Her work in value sensitive design has resulted in robust theoretical constructs, dozens of innovative methods, and practical toolkits such as the Envisioning Cards. Value sensitive design has been widely adopted nationally and internationally where it has been used in architecture, biomedical health informatics, civil engineering, computer security, energy, global health, human-computer interaction, human-robotic interaction, information management, legal theory, moral philosophy, tech policy, transportation, and urban planning, among others. Additionally, value sensitive design is emerging in higher education, government, and industry as a key approach to address computing ethics and responsible innovation. Today, Dr. Friedman is working on open questions in value sensitive design including multi-lifespan design, and designing for and with non-human stakeholders – questions critical for the wellbeing of human societies and the planet.

Dr. Friedman's 2019 MIT Press book co-authored with David Hendry, Value Sensitive Design: Shaping Technology with Moral Imagination, provides a comprehensive account of value sensitive design. In 2012 Dr. Friedman received the ACM-SIGCHI Social Impact Award and the University Faculty Lecturer award at the University of Washington, in 2019 she was inducted into the CHI Academy, in 2020 she received an honorary doctorate from Delft University of Technology, and in 2021 she was recognized as an ACM Fellow. She is also a stone sculptor and mixed media artist. Dr. Friedman received both her B.A. and Ph.D. from the University of California at Berkeley.

NLP in Mexican Spanish: One of many stories

Manuel Montes-y-Gómez
National Institute of Astrophysics, Optics and Electronics (INAOE)

Wednesday, July 13, 2022 - Room: **Columbia A/C/D & 302 Beckler** - Time: **16:15-17:15**

Abstract: Spanish is one of the most widely spoken languages in the world, however, the development of language technologies for it has not been in the same proportion. This is particularly true for some of its Latin American variants, such as the Mexican Spanish. This talk will focus on presenting the development of NLP for Mexican Spanish, emphasizing one of its many research stories related to the analysis of social media content.

This talk will present some data on the languages spoken in Mexico and on the development of the area of Natural Language Processing in our country, and will describe a research project that combined the efforts of several groups: the identification of abusive language in Mexican tweets. The talk will conclude by exposing some calls for collaboration, with the intention of increasing and improving the research in Mexican Spanish as well as in the many indigenous languages spoken in Mexico.

Bio: Manuel Montes-y-Gómez is Full Professor at the National Institute of Astrophysics, Optics and Electronics (INAOE) of Mexico. His research is on automatic text processing. He is author of more than 250 journal and conference papers in the fields of information retrieval, text mining and authorship analysis.

He has been visiting professor at the Polytechnic University of Valencia (Spain), and the University of Alabama (USA). He is also a member of the Mexican Academy of Sciences (AMC), and founding member of the Mexican Academy of Computer Science (AMEXCOMP), the Mexican Association of Natural Language Processing (AMNLP), and of the Language Technology Network of CONACYT. In the context of them, he has been the organizer of the National Workshop on Lanuage Technologies (from 2004 to 2016), the Mexican Workshop on Plagiarism Detection and Authorship Analysis (2016-2020), the Mexican Autumn School on Language Technologies (2015 and 2016), and a shared task on author profiling, aggressiveness analysis and fake news detection in Mexican Spanish at IberLEF (2018-2021).

Panel: "The Place of Linguistics and Symbolic Structures"

Tuesday, July 12, 2022 - Room: **Columbia A/C/D & 302 Beckler** - Time: **9:15-10:15**

The widespread adoption of neural models in NLP research and the fact that NLP applications increasingly mediate people's lives have prompted many discussions about what productive research directions might look like for our community. Since NAACL is a meeting of a chapter of the Association for Computational Linguistics, we would like to highlight specifically the role that linguistics and symbolic structures can play (or not) in shaping these research directions.

Moderator: Dan Roth, University of Pennsylvania & AWS AI Labs
Bio: Dan Roth is the Eduardo D. Glandt Distinguished Professor at the Department of CIS, UPenn, the NLP Lead at AWS AI, and a Fellow of the AAAS, ACM, AAAI, and ACL. In 2017 Roth received the John McCarthy Award. Roth has published broadly in ML, NLP, KRR, and learning theory, and has given keynote talks and tutorials in all ACL and AAAI major conferences. Roth was the Editor-in-Chief of JAIR until 2017, and the program chair of AAAI'11, ACL'03 and CoNLL'02.

Emily M. Bender, University of Washington
Bio: Emily M. Bender is a Professor of Linguistics at the University of Washington and the Faculty Director of UW's Professional Master's in Computational Linguistics. Her research interests include computational semantics, multilingual grammar engineering, the interplay between linguistics and NLP, and societal impacts of language technology. She is the author of two books which present linguistic concepts in a manner accessible to NLP practitioners: Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax (2013) and Linguistic Fundamentals for Natural Language Processing II: 100 Essentials from Semantics and Pragmatics (2019; with Alex Lascarides), as well as the co-author of recent influential papers such as Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data (ACL 2020) and On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? (FAcct 2021).

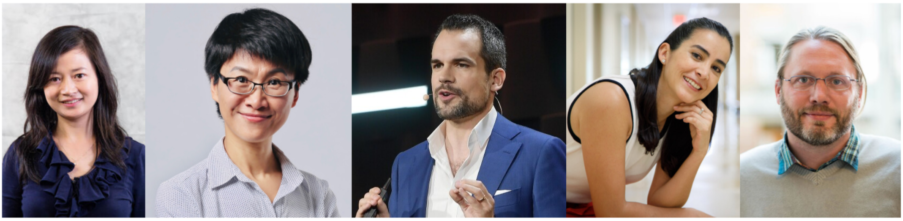
Dilek Hakkani-Tür, Amazon Alexa AI
Bio: Dilek Hakkani-Tür is a senior principal scientist at Amazon Alexa AI, focusing on enabling natural dialogues with machines. Prior to joining Amazon, she was a researcher at Google, Microsoft Research, International Computer Science Institute at UC Berkeley and AT&T Labs-Research. Her research interests include conversational AI, natural language and speech processing, spoken dialogue systems, and machine learning for language processing. She received best paper awards for publications she co-authored on conversational systems from IEEE Signal Processing Society, ISCA and EURASIP. Recently, she served as a program chair for NAACL 2020, the editor-in-chief of IEEE Transactions on Audio, Speech, and Language Processing and an IEEE Distinguished Industry Speaker. She is a fellow of ISCA and IEEE.

Chitta Baral, Arizona State University
Bio: Chitta Baral is a Professor in the School of Computing and AI at Arizona State University. His

research interests include Knowledge Representation and Reasoning (KR & R), Natural Language Understanding (NLU), Image/Video Understanding; and their applications to Molecular Biology, Health Informatics and Robotics. Chitta is the author of the book "Knowledge Representation, Reasoning and Declarative Problem Solving" and a past President of KR Inc. His current research focus is on leveraging decades of research in KR & R for better understanding of natural language and images/videos. Towards that end he has worked on a framework for translating natural language to formal representations (NL2KR); abducing missing knowledge and knowledge hunting; exploring NLU challenges where reasoning with knowledge, reasoning about actions, and commonsense reasoning are crucial; exploring the use of natural language as a knowledge representation and instructional formalism; and exploring the role of reasoning and knowledge in enhancing generalizability, robustness, and few-shot learning.

Christopher D. Manning, Stanford University
Bio: Christopher Manning is a professor of linguistics and computer science at Stanford University, Director of the Stanford Artificial Intelligence Lab (SAIL), and an Associate Director of the Stanford Institute for Human-Centered AI (HAI). He is a leader in applying deep neural networks to natural language processing (NLP), including work on neural machine translation, tree-recursive models, natural language inference, summarization, parsing, question answering, and the GloVe word vectors. Manning founded the Stanford NLP group (@stanfordnlp), teaches and has co-written textbooks for NLP (CS 224N) and information retrieval (CS 276), co-developed Stanford Dependencies and Universal Dependencies, manages development of the Stanford CoreNLP and Stanza software, is the most-cited researcher in NLP, and is an ACM, AAAI, and ACL Fellow and a Past President of ACL.

Panel: “Careers in NLP”



Monday, July 11, 2022 - Room: **Columbia A/C/D & 302 Beckler** - Time: **13:15-14:15**

The Careers in NLP Panel is a standing feature of NAACL Industry Track. The panel is addressed to graduate students and junior researchers as well as their supervisors and mentors, although all NAACL participants are welcomed. The panellists will discuss the diversity of career paths in NLP: from more research-oriented NLP scientist roles to careers in product.

Moderator: Yunyao Li, Apple Knowledge Platform
Bio: Yunyao Li is the Head of Machine Learning, Apple Knowledge Platform, where her team builds the next-generation machine learning solutions to help power features such as Siri and Spotlight. Previously she was a Distinguished Research Staff Member and Senior Research Manager at IBM Research - Almaden. She is particularly known for her work in scalable NLP, enterprise search, and database usability. She has built systems, developed solutions, and delivered core technologies to over 20 IBM products under brands such as Watson, InfoSphere, and Cognos. She has published over 80 articles with multiple awards and a book. She was an IBM Master Inventor, with over 50 patents filed/granted. She is an ACM Distinguished Member. She was a member of the inaugural New Voices program of the US National Academies (1 out of 18 selected nationwide) and represented US young scientists at World Laureates Forum Young Scientists Forum in 2019 (1 of 4 selected nationwide).

Yang Liu, Amazon, Alexa AI
Bio: Yang Liu is currently a principal scientist at Amazon, Alexa AI. Her research interest is in speech and language processing. She received her BS and MS from Tsinghua University, and Ph.D. from Purdue University. Before joining Amazon, she was the head of LAIX Silicon Valley AI lab, a research scientist at Facebook, visiting scientist at Google, a faculty member at the University of Texas at Dallas, and researcher at ICSI in Berkeley. She received NSF CAREER award and Air Force Young Investigator Program award. She is currently a member of the IEEE SLTC committee, a senior area editor for IEEE/ACM Transactions on Audio, Speech and Language Processing, an action editor for TACL. She was one of the program chairs for EMNLP 2020, and has served regularly as an area chair and reviewer in the past NLP conferences. She is a fellow of IEEE and ISCA.

Timo Mertens, Grammarly
Bio: Timo Mertens is the Head of Machine Learning & NLP Products at Grammarly. In his role, he oversees the teams that design and build products that use machine learning and natural language processing. These technologies empower Grammarly to offer a digital writing assistant that helps millions of users write more clearly and effectively every day. Timo has focused on the intersection between machine learning and delivering impactful products throughout his career, spanning academia—with a Ph.D. in Speech Recognition—and industry, where he’s held product leadership positions across Microsoft, Google, and Dropbox.

Thamar Solorio, University of Houston and Bloomberg LP
Bio: Thamar Solorio is a Professor of Computer Science at the University of Houston (UH) and she is also a visiting scientist at Bloomberg LP. She holds graduate degrees in Computer Science from the Instituto Nacional de Astrofísica, Óptica y Electrónica, in Puebla, Mexico. Her research interests include information extraction from social media data, enabling technology for code-switched data, stylistic modelling of text, and more recently multimodal approaches for online content understanding. She is the director and founder of the Research in Text Understanding and Language Analysis Lab at UH. She is the recipient of an NSF CAREER award for her work on authorship attribution, and recipient of the 2014 Emerging Leader ABIE Award in Honor of Denice Denton. She is currently serving a second term as an elected board member of the North American Chapter of the Association of Computational Linguistics.

Luke Zettlemoyer, University of Washington and Meta
Bio: Luke Zettlemoyer is a Professor in the Paul G. Allen School of Computer Science & Engineering at the University of Washington, and a Research Scientist at Meta. His research focuses on empirical methods for natural language semantics, and involves designing machine learning algorithms, introducing new tasks and datasets, and, most recently, studying how to best develop self-supervision signals for pretraining. His honors include being named an ACL Fellow as well as winning a PECASE award, an Allen Distinguished Investigator award, and multiple best paper awards.

6

D&I Events and Initiatives

Affinity Groups

For NAACL 2022, the D&I committee has taken several steps to strengthen the presence and bolster visibility of affinity groups in NLP. We intend this guide to be a primer for those who have not heard about them or are interested in better supporting their efforts.

What is an Affinity Group?

An affinity group is a collective of researchers formed around a shared interest or common goal.

For example, they might be comprised of individuals from specific protected categories or are from traditionally underrepresented geographic regions. In our community, affinity groups promote and support ideas and voices of underrepresented groups and raise awareness of issues that affect their members.

What Kind of Events do Affinity Group Offer at Conferences?

Affinity groups organize social events for their members at conferences, which the D&I committee provides support for. NAACL 2022 will additionally feature several D&I panels where affinity groups members will participate towards the shared goal for supporting diversity, equality and inclusion for all.

What Can NAACL Attendees do to Support Affinity Groups?

1. Participate - Actively attending and participating in panels and sessions for discussions about issues that are relevant to the affinity groups goes a long way in helping promote their cause. We request all members of our community to participate actively!

2. Sponsor, Volunteer or Recruit - Affinity groups are always looking for help with sponsorship, volunteering and recruitment. Help with sponsorship, volunteer your time, and help recruitment efforts to keep the momentum going!

For the first time in *CL conferences, we are organizing affinity group workshops.

Queer in AI

Queer in AI was established by queer scientists in AI with the mission to make the AI community a safe and inclusive place that welcomes, supports, and values queer people. We work towards this aim by building a visible community of queer AI scientists through conference workshops, social meetups, conference poster sessions, mentoring programs, graduate application financial aid, and many other initiatives. Another crucial part of our mission is to raise awareness of queer issues in the general AI community and to encourage and highlight research on and solutions to these problems.

You can apply for volunteering by participating in their slack. They have many different roles like helping with conference socials, mentoring people through grad school applications or curating our twitter account

and youtube channel. They are especially looking for people to help with finance and sponsorship, and help organize our presence at future NLP conferences.
Workshop details:
Queer in AI program includes the following:

1. Virtual social on July 9, 7 am PST

2. In-person workshop at 506 Samish on July 10, 9 am PST

More information will be available on Queer in AI website, `www.queerinai.com/naacl`.
List of accepted papers:

1. Detecting Harmful Online Conversational Content Detection towards LGBTQIA Individuals. *Jamell Dacon, Harry Shomer, Shaylynn L.A. Crum-Dacon, Jiliang Tang*

2. Outed by an Algorithm: A Study on Facebook's Friends Recommendation System for Queer, Trans and Gender-Non-Conforming Users. *Jack Chang*

3. Towards WinoQueer: Developing a Benchmark for Anti-Queer Bias in Large Language Models. *Virginia K. Felkner, Ho-Chun Herbert Chang, Eugene Jang, Jonathan May*

4. Use of a Stylometric Map-Based Corpus for Tracking Individual Variation in Relation to Gender and Sex. *Theodore Daniel Manning, Harleigh Niyu, Alejandro Jorge Napolitano Jawerbaum, Patrick Juola*

5. Overview of STEM Science as Process, Data, Material, and Method Named Entities. *Jennifer D'Souza*

6. CNSC: Czech news article dataset for classification of originating source and its credibility. *Matyáš Bohácek*

7. Tackling Gender Microaggressions in Hindi. *Vishakha Agrawal*

Queer in AI code of conduct will be strictly followed at all times at the Queer in AI events. Recording (screen recording or screenshots) is prohibited. All participants are expected to maintain the confidentiality of other participants.

LatinX in AI

The LatinX in AI (LXAI) network serves many researchers around the globe. This time, LXAI organized the first NLP workshop aimed at LatinX to promote and increase the representation of the LatinX community in the NLP field. The LXNLP workshop includes paper presentations, a diverse set of keynotes (speakers come from academia, government, and industry), a panel discussion, and a social. For more information, please visit the LXNLP website: `latinxinai.org/naacl-2022`.
The in-person workshop for LatinX in AI will be on July 10, 11.30 am PST at 507 Sauk.
List of accepted papers:

1. Incorporating Natural Language Processing models in Mexico City's 311 Locatel *Alejandro Molina-Villegas, Edwin Aldana-Bibadilla, Oscar S Siordia, Jorge Luis Perez*

2. An interpretable representation of dialog history in referential visual dialog *Mauricio Mazuecos, Franco Luque, Jorge Sánchez, Hernán Maina, Thomas Vadora, Luciana Benotti*

3. Identifying epidemic related Tweets using noisy learning *Ramya Tekumalla, Juan M. Banda*

4. Automatic multi-modal processing of language and vision to assist people with visual impairments *Hernán Maina, Luciana Benotti*

5. Distributed Text Representations Using Transformers for Noisy Written Language *Alejandro Rodriguez Perez, Pablo Rivas Perea, Gissella Bejarano Nicho*

6. BioMedIA: A Complete Voice-to-Voice Generative Question Answering System for the Biomedical Domain in Spanish *Alejandro Vaca Serrano, David Betancur Sánchez, Alba Segurado, Guillem García Subies, Álvaro Barbero Jiménez*

7. Dual Architecture for Name Entity Extraction and Relation Extraction with Applications in Medical Corpora *Ernesto Quevedo Caballero, Alejandro Rodriguez Perez, Tomas Cerny, Pablo Rivas*

8. User Profile Characterization Within a Brazilian Online Dispute Resolution Platform *Wesley Paulino Fernandes Maciel, Yohan Bonescki Gumiel, Adriana Pagano, Ana Paula Couto da Silva*

9. Study of Question Answering on Legal Software Document using BERT based models *Ernesto Quevedo Caballero, Mushfika Rahman, Tomas Cerny, Pablo Rivas, Gissella Bejarano*

10. Improving Language Model Fine-tuning with Information Gain Filtration *Javier S. Turek, Richard Antonello, Nicole M. Beckage, Alexander G. Huth*

North Africans in NLP

The goal of this group is to create a sense of community among North African researchers in NLP & AI, to increase the visibility of North Africans within the NLProc community, to highlight their accomplishments, and to acknowledge and discuss the hardships faced by NLP researchers based in North Africa. There will be talks about work on North African languages (Tamazight and North African Arabic dialects). All are welcome!
This is the virtual only workshop. More details are available at `https://sites.google.com/view/northafricansinnlp/naacl-2022`
List of accepted papers:

1. Tunisian Dialectal Speech Recognition Model by *Abir Messaoudi and Hatem Haddad.*

2. GOUD.MA: A News Article Dataset For Summarization In Moroccan Darija by *Abderrahmane Issam and Khalil Mrini*

Widening NLP

WiNLP started as a general cohort building, to form networks of like-minded people. They want to find the right vectors of support for more and different groups within CL. One size does not fit all when it comes to diversity and support. They wanted to create a space that would be welcoming and supportive for all groups, particularly those that typically don't get to see themselves represented as often or well in the larger CL community.
If you would like to help WiNLP, then volunteer and sponsor! Reach out to an organizer to say you want to be on the organizing committee next year; volunteer to help run things on the day of as an extra pair of hands; share the calls for papers and other WiNLP events; encourage your students and labs to submit and attend our workshops; sign up as a mentor. Follow WiNLP goals outside the workshop itself — better representation of underrepresented minorities in panel discussions, for invited talks, ACL fellowships, etc.
Workshop details
This is a virtual only workshop which will take place on July 10. More details are available at their website: `http://www.winlp.org/winlp-satellite-workshop-naacl-2022/`

1. Intro and Keynote (8-9 am PDT)
2. Mentorship Lunch (12-1 pm PDT)
3. Panel and goodbye (5-6 pm PDT)

Acknowledgement

The NAACL DEI committee and the NAACL Student Research Workshop (SRW) would like to express gratitude to our wonderful mentors for volunteering for our pre-submission mentoring programs:

Valerio Basile, Kasturi Bhattacharjee, Yonatan Bisk, Eduardo Blanco, Muhao Chen, Anna Currey, Sunipa Dev, Greg Durrett, Dan Goldwasser, Kristen Johnson, Mihir Kale, Mascha Kurpicz-Briki, Sandeep Mathias, Paul Michel, Shachar Mirkin, Sheshera Mysore, Vincent Ng, Atul Kumar Ojha, Nedjma Ousidhoum, Archita Pathak, Rajaswa Patil, Melissa Roemmele, Emily Sheng, Vivek Srivastava, Hao Tan, David Tresner-Kirsch, Shomir Wilson, Arkaitz Zubiaga

Their support made it possible for us to offer new researchers, including first-time submitters, the opportunity to receive early feedback on their submissions from more experienced researchers. Thank you for helping us make NAACL more welcoming for these researchers!

7

Tutorials: Sunday, July 10, 2022

Overview

07:30 - 18:00	***Registration***	
08:00 - 08:45	***Extra Q&A 1 - Morning Tutorials***	
	Tutorial 1 – Text Generation with Text-Editing Models Eric Malmi, Yue Dong, Jonathan Mallinson, Aleksandr Chuklin, Jakub Adamek, Daniil Mirylenka, Felix Stahlberg, Sebastian Krause, Shankar Kumar, Aliaksei Severyn	*Ballroom Columbia A*
	Tutorial 2 – Self-supervised Representation Learning for Speech Processing Hung-yi Lee, Abdelrahman Mohamed, Shinji Watanabe, Tara Sainath, Karen Livescu, Shang-Wen Li, Shu-wen Yang, Katrin Kirchhoff	*Ballroom Columbia C*
	Tutorial 3 –New Frontiers of Information Extraction Muhao Chen, Lifu Huang, Manling Li, Ben Zhou, Heng Ji, Dan Roth	*Ballroom Columbia D*
09:00 - 12:30	***Morning Tutorials***	
	Tutorial 1 – Text Generation with Text-Editing Models Eric Malmi, Yue Dong, Jonathan Mallinson, Aleksandr Chuklin, Jakub Adamek, Daniil Mirylenka, Felix Stahlberg, Sebastian Krause, Shankar Kumar, Aliaksei Severyn	*Ballroom Columbia A*
	Tutorial 2 – Self-supervised Representation Learning for Speech Processing Hung-yi Lee, Abdelrahman Mohamed, Shinji Watanabe, Tara Sainath, Karen Livescu, Shang-Wen Li, Shu-wen Yang, Katrin Kirchhoff	*Ballroom Columbia C*

Tutorial 3 –New Frontiers of Information Extraction *Ballroom Columbia D*

Muhao Chen, Lifu Huang, Manling Li, Ben Zhou, Heng Ji, Dan Roth

12:30 - 13:00 ***Extra Q&A 2 - Morning Tutorials***

Tutorial 1 – Text Generation with Text-Editing Models *Ballroom Columbia A*

Eric Malmi, Yue Dong, Jonathan Mallinson, Aleksandr Chuklin, Jakub Adamek, Daniil Mirylenka, Felix Stahlberg, Sebastian Krause, Shankar Kumar, Aliaksei Severyn

Tutorial 2 – Self-supervised Representation Learning for Speech Processing *Ballroom Columbia C*

Hung-yi Lee, Abdelrahman Mohamed, Shinji Watanabe, Tara Sainath, Karen Livescu, Shang-Wen Li, Shu-wen Yang, Katrin Kirchhoff

Tutorial 3 –New Frontiers of Information Extraction *Ballroom Columbia D*

Muhao Chen, Lifu Huang, Manling Li, Ben Zhou, Heng Ji, Dan Roth

13:30 - 14:00 ***Extra Q&A 1 - Afternoon Tutorials***

Tutorial 4 – Human-Centered Evaluation of Explanations *Ballroom Columbia D*

Jordan Boyd-Graber, Samuel Carton, Shi Feng, Vera Liao, Tania Lombrozo, Alison Smith-Renner, Chenhao Tan

Tutorial 5 – Multimodal Machine Learning *Ballroom Columbia C*

Louis-Philippe Morency, Paul Pu Liang, Amir Zadeh

Tutorial 6 – Contrastive Data and Learning for Natural Language Processing *Ballroom Columbia A*

Rui Zhang, Yangfeng Ji, Yue Zhang, Rebecca J. Passonneau

14:00 - 17:30 ***Afternoon tutorials***

Tutorial 4 – Human-Centered Evaluation of Explanations *Ballroom Columbia D*

Jordan Boyd-Graber, Samuel Carton, Shi Feng, Vera Liao, Tania Lombrozo, Alison Smith-Renner, Chenhao Tan

Tutorial 5 – Multimodal Machine Learning *Ballroom Columbia C*

Louis-Philippe Morency, Paul Pu Liang, Amir Zadeh

Tutorial 6 – Contrastive Data and Learning for Natural Language Processing *Ballroom Columbia A*

Rui Zhang, Yangfeng Ji, Yue Zhang, Rebecca J. Passonneau

18:00 - 18:45 ***Extra Q&A 2 - Afternon Tutorials***

Tutorial 4 – Human-Centered Evaluation of Explanations *Ballroom Columbia D*

Jordan Boyd-Graber, Samuel Carton, Shi Feng, Vera Liao, Tania Lombrozo, Alison Smith-Renner, Chenhao Tan

Tutorial 5 – Multimodal Machine Learning *Ballroom Columbia C*
Louis-Philippe Morency, Paul Pu Liang, Amir Zadeh

Tutorial 6 – Contrastive Data and Learning for Natural Language Processing *Ballroom Columbia A*
Rui Zhang, Yangfeng Ji, Yue Zhang, Rebecca J. Passonneau

Welcome to the Tutorials Session of NAACL 2022!

The tutorials give an opportunity to the NAACL conference attendees to be lectured by highly qualified expert researchers on cutting-edge and new relevant upcoming topics in our research community.

As in previous years, the organization (including submission, reviewing and selection) were coordinated jointly with other conferences in the 2022 calendar year: ACL, NAACL, COLING and EMNLP. We formed a review committee of 34 members, which includes the NAACL tutorial chairs, the ACL tutorial chairs, the COLING tutorial chairs, the EMNLP tutorial chairs and 23 external reviewers (see Program Committee for the full list). We organized a reviewing process so that each proposal received at least 3 reviews. Tutorials were evaluated based on their clarity, novelty, timely character of the topic, diversity and inclusion, instructor's experience, likely audience interest and open access of the tutorial instructional material. We received a total of 47 tutorial submissions, of which 6 were selected for presentation at NAACL, considering the preferences expressed by authors and the relevance for the NAACL research community.

We solicited two types of tutorials, namely cutting-edge themes and introductory themes. The 6 tutorials for NAACL include one introductory tutorial and five cutting-edge tutorials. The introductory tutorial is dedicated to Human-Centered Evaluation of Explanations (T4). The cutting-edge tutorials are: (T1) Text Generation with Text-Editing Models, (T2) Self-supervised Representation Learning for Speech Processing, (T3) New Frontiers of Information Extraction, (T5) Multimodal Machine Learning, and (T6) Contrastive Data and Learning for Natural Language Processing. NAACL 2022 tutorials are delivered in a live hybrid format and also available as pre-recorded captioned videos, with additional live Q&A sessions.

We would like to thank the tutorial authors for their quick responses and flexibility while organizing the conference in a hybrid mode. We are also grateful to the 23 external reviewers for their invaluable help in the decision process. Finally, we thank the conference organizers for effective collaboration, the general chair Dan Roth, the program chairs (Marine Carpuat, Marie-Catherine de Marneffe and Ivan Vladimir Meza Ruiz), the publication chair Ryan Cotterell, and the authors of `aclpub2` with special mention to Jordan Zhang and Danilo Croce.

NAACL 2022 Tutorial Co-chairs,

Miguel Ballesteros Yulia Tsvetkov Cecilia O. Alm

T1 - Text Generation with Text-Editing Models

Eric Malmi, Yue Dong, Jonathan Mallinson, Aleksandr Chuklin, Jakub Adamek, Daniil Mirylenka, Felix Stahlberg, Sebastian Krause, Shankar Kumar, Aliaksei Severyn

Cutting-edge

Sunday, July 10, 2022 - 9:00-12:30. Extra Q&A Sessions: 8:00-8:45 and 12:30-13:00 (Ballroom Columbia A)

`https://text-editing.github.io/`

Text-editing models have recently become a prominent alternative to seq2seq models for monolingual text-generation tasks such as grammatical error correction, text simplification, and style transfer. These tasks share a common trait – they exhibit a large amount of textual overlap between the source and target texts. Text-editing models take advantage of this observation and learn to generate the output by predicting edit operations applied to the source sequence. In contrast, seq2seq models generate outputs word-by-word from scratch thus making them slow at inference time. Text-editing models provide several benefits over seq2seq models including faster inference speed, higher sample efficiency, and better control and interpretability of the outputs. This tutorial provides a comprehensive overview of the text-edit based models and current state-of-the-art approaches analyzing their pros and cons. We discuss challenges related to deployment and how these models help to mitigate hallucination and bias, both pressing challenges in the field of text generation.

Eric Malmi, Google, Switzerland
Website: `https://ericmalmi.com/`
Eric Malmi is a Senior Research Scientist at Google Switzerland. His research is focused on developing text-generation models for grammatical error correction and text style transfer. He received his PhD from Aalto University, Finland.

Yue Dong, McGill University and Mila, Canada
Website: `https://www.cs.mcgill.ca/~ydong26/`
Yue Dong is a final-year PhD student in CS at McGill University and Mila, Canada. Her research is focused on conditional text generation. She is a co-organizer for the NewSum workshop at EMNLP 2021 and ENLSP workshop at NeurIPS 2021.

Jonathan Mallinson, Google, Switzerland
Jonathan Mallinson is a Research Engineer at Google Switzerland. His research is focused on low-latency text-to-text generation. He received his PhD from the University of Edinburgh, Scotland.

Aleksandr Chuklin, Google, Switzerland
Website: `http://linkedin.com/in/chuklin/`
Aleksandr Chuklin is a Research Engineer at Google Switzerland. His current research focuses on multilingual NLG. He organized workshops and conducted tutorials at conferences such as SIGIR, EMNLP, and IJCAI. Aleksandr received his PhD from University of Amsterdam, The Netherlands.

Jakub Adamek, Google, Switzerland
Website: `http://linkedin.com/in/jakub-adamek-pl/`
Jakub Adamek is a Research Engineer at Google Switzerland focusing on grammatical error correction and low-latency models. Jakub received his MSc from Jagiellonian University, Poland.

Daniil Mirylenka, Google, Switzerland
Website: `http://linkedin.com/in/daniil-mirylenka-b0428a26`
Daniil Mirylenka is a Research Engineer at Google Switzerland working on text editing with application to grammatical error correction. Daniil received his PhD from the University of Trento, Italy.

Felix Stahlberg, Google, Switzerland
Felix Stahlberg is a Research Scientist at Google focusing on grammatical error correction and text style models. Felix received his PhD from Cambridge University, UK.

Sebastian Krause, Google, Switzerland
Website: `https://scholar.google.com/citations?user=i03iu-UAAAAJ&hl=en`
Sebastian Krause is a Senior Research Engineer at Google Switzerland. His work is focused on multilingual rewriting of questions in low-latency settings. Sebastian received his PhD in Engineering from the Technical University of Berlin, Germany.

Shankar Kumar, Google, Switzerland
Shankar Kumar is a Senior Staff Research Scientist at Google leading a research team working on speech and language algorithms. He received his PhD from the Johns Hopkins University, US.

Aliaksei Severyn, Google, Switzerland
Website: `http://linkedin.com/in/aseveryn`
Aliaksei Severyn is a Staff Research Scientist at Google Switzerland leading an applied research team working on next generation NLG solutions. Aliaksei received his PhD from University of Trento, Italy.

T2 - Self-Supervised Representation Learning for Speech Processing

Hung-yi Lee, Abdelrahman Mohamed, Shinji Watanabe, Tara Sainath, Karen Livescu, Shang-Wen Li, Shu-wen Yang, Katrin Kirchhoff

Cutting-edge

Sunday, July 10, 2022 - 9:00-12:30. Extra Q&A Sessions: 8:00-8:45 and 12:30-13:00 (Ballroom Columbia C)

`https://sites.google.com/view/tutorial-ssl-speech`

Although Deep Learning models have revolutionized the speech and audio processing field, they forced building specialist models for individual tasks and application scenarios. Deep neural models also bottlenecked dialects and languages with limited labeled data. Self-supervised representation learning methods promise a single universal model to benefit a collection of tasks and domains. They recently succeeded in NLP and computer vision domains, reaching new performance levels while reducing required labels for many downstream scenarios. Speech representation learning is experiencing similar progress with three main categories: generative, contrastive, predictive. Other approaches relied on multi-modal data for pre-training, mixing text or visual data streams with speech. Although self-supervised speech representation is still a nascent research area, it is closely related to acoustic word embedding and learning with zero lexical resources. This tutorial session will present self-supervised speech representation learning approaches and their connection to related research areas. Since many of the current methods focused solely on automatic speech recognition as a downstream task, we will review recent efforts on benchmarking learned representations to extend the application of such representations beyond speech recognition. A hands-on component of this tutorial will provide practical guidance on building and evaluating speech representation models.

Hung-yi Lee, National Taiwan University
Website: `https://speech.ee.ntu.edu.tw/~hylee/index.php`
Hung-yi Lee received the Ph.D. degree from National Taiwan University (NTU). He was a visiting scientist at the Spoken Language Systems Group of MIT CSAIL. He is an associate professor at National Taiwan University. He is the co-organizer of the special session on “New Trends in self-supervised speech processing” at Interspeech (2020), and the workshop on "Self-Supervised Learning for Speech and Audio Processing" at NeurIPS (2020).

Abdelrahman Mohamed, Meta AI, USA
Website: `https://ai.facebook.com/people/abdelrahman-mohamed/`
Abdelrahman Mohamed is a research scientist at Meta AI research. He received his PhD from the University of Toronto where he was part of the team that started the Deep Learning revolution in Spoken Language Processing in 2009. He has been focusing lately on improving, using, and benchmarking learned speech

representations, e.g. HuBERT, Wav2vec 2.0, TextlessNLP, and SUPERB.

Shinji Watanabe, Carnegie Mellon University, Pittsburgh, USA.
Website: `https://sites.google.com/view/shinjiwatanabe`
Shinji Watanabe is an Associate Professor at CMU. He was a research scientist at NTT, Japan, a visiting scholar in Georgia Tech, a senior principal research scientist at MERL, and an associate research professor at JHU. He has published more than 200 peer-reviewed papers. He served as an Associate Editor of the IEEE TASLP. He was/has been a member of several technical committees, including the APSIPA SLA, IEEE SPS SLTC, and MLSP.

Tara Sainath, Google, New York, USA
Website: `https://research.google/people/TaraSainath/`
Tara Sainath is a Principal Research scientist at Google. She received her PhD from MIT in the Spoken Language Systems Group. She is an IEEE and ISCA Fellow and the recipient of the 2021 IEEE SPS Industrial Innovation Award. Her research involves applications of deep neural networks for automatic speech recognition, and has been very active in the community organizing workshops and special sessions on this topic.

Karen Livescu, Toyota Technological Institute at Chicago (TTIC), USA
Website: `https://ttic.edu/livescu/`
Karen Livescu is a Professor at TTI-Chicago. She completed her PhD at MIT in the Spoken Language Systems group. She is an ISCA Fellow and an IEEE Distinguished Lecturer, and has served as a program chair for ICLR 2019 and Interspeech 2022. Her recent work includes multi-view representation learning, acoustic word embeddings, visually grounded speech models, spoken language understanding, and automatic sign language recognition.

Shang-Wen Li, Meta AI
Website: `https://swdanielli.github.io/`
Shang-Wen Li is a Research and Engineering Manager at Meta AI. He worked at Apple Siri, Amazon Alexa and AWS. He completed his PhD in 2016 from the Spoken Language Systems group of MIT CSAIL. He co-organized the workshop of "Self-Supervised Learning for Speech and Audio Processing" at NeurIPS (2020) and AAAI (2022), and the workshop of "Meta Learning and Its Applications to Natural Language Processing" at ACL (2021).

Shu-wen Yang, National Taiwan University
Website: `https://scholar.google.com.tw/citations?user=R1mNI8QAAAAJ`
Shu-wen Yang is a Ph.D. student at National Taiwan University. He co-created a benchmark for Self-Supervised Learning in speech, Speech processing Universal PERformance Benchmark (SUPERB). Before SUPERB, he created the S3PRL toolkit with Andy T. Liu, which supports numerous pretrained models and recipes for both pre-training and benchmarking. He gave a tutorial at the Machine Learning Summer School, Taiwan, 2021.

Katrin Kirchhoff, AWS AI Labs, Seattle, USA
Website: `https://www.amazon.science/author/katrin-kirchhoff`
Katrin is a Director of Applied Science at Amazon Web Services, where she heads several teams in speech and audio processing. She was a Research Professor at the UW, Seattle, for 17 years, where she co-founded the Signal, Speech and Language Interpretation Lab. She served on the editorial boards of Speech Communication and Computer, Speech, and Language, and was a member of the IEEE Speech Technical Committee.

T3 - New Frontiers of Information Extraction

Muhao Chen, Lifu Huang, Manling Li, Ben Zhou, Heng Ji, Dan Roth

Cutting-edge

Sunday, July 10, 2022 - 9:00-12:30. Extra Q&A Sessions: 8:00-8:45 and 12:30-13:00 (Ballroom Columbia D)

https://cogcomp.seas.upenn.edu/page/tutorial.202207/

Information extraction (IE) is the process of automatically extracting structural information from unstructured or semi-structured data. It provides the essential support for natural language understanding by recognizing and resolving the concepts, entities, events described in text, and inferring the relations among them. In various application domains, IE automates the costly acquisition process of domain-specific knowledge representations that have been the backbone of any knowledge-driven AI systems. For example, automated knowledge base construction has relied on technologies for entity-centric IE. Extraction of events and event chains assists machines with narrative prediction and summarization tasks. Medical IE also benefits important but expensive clinical tasks such as drug discovery and repurposing. Despite the importance, frontier research in IE still face several key challenges. The first challenge is that existing dominant methods using language modeling representation cannot sufficiently capture the essential knowledge and structures required for IE tasks. The second challenge is on the development of extraction models for fine-grained information with less supervision, considering that obtaining structural annotation on unlabeled data have been very costly. The third challenge is to extend the reliability and generalizability of IE systems in real-world scenarios, where data sources often contain incorrect, invalid or unrecognizable inputs, as well as inputs containing unseen labels and mixture of modalities. Recently, by tackling those critical challenges, recent literature is leading to transformative advancement in principles and methodologies of IE system development. We believe it is necessary to present a timely tutorial to comprehensively summarize the new frontiers in IE research and point out the emerging challenges that deserve further investigation.

In this tutorial, we will systematically review several lines of frontier research on developing robust, reliable and adaptive learning systems for extracting rich structured information. Beyond introducing robust learning and inference methods for unsupervised denoising, constraint capture and novelty detection, we will discuss recent approaches for leveraging indirect supervision from natural language inference and generation tasks to improve IE. We will also review recent minimally supervised method for training IE models with distant supervision from linguistic patterns, corpus statistics or language modeling objectives. In addition, we will illustrate how a model trained on a close domain can be reliably adapted to produce extraction from data sources in different domains, languages and modalities, or acquiring global knowledge to guide the extraction on a highly diverse open label space. Participants will learn about recent trends and emerging challenges in this topic, representative tools and learning resources to obtain ready-to-use models, and how related technologies benefit end-user NLP applications.

Muhao Chen, University of Southern California, USA
Website: `https://luka-group.github.io/`
Muhao Chen is an Assistant Research Professor of Computer Science at USC, where he directs the USC Language Understanding and Knowledge Acquisition (LUKA) Group. His research focuses on data-driven machine learning approaches for natural language understanding and knowledge acquisition. His work has been recognized with an NSF CRII Award, an ACM SIGBio Best Student Paper Award, and a Best Paper Nomination at CoNLL.

Lifu Huang, Virginia Tech
Website: `https://wilburone.github.io/`
Lifu Huang is an Assistant Professor at the Computer Science Department of Virginia Tech. He has a wide range of research interests in NLP, including extracting structured knowledge with limited supervision, NLU and reasoning with external knowledge and commonsense, NLG, representation learning for cross-lingual and cross-domain transfer, and multi-modality learning. He is a recipient of the 2019 AI2 Fellowship and 2021 Amazon Research Award.

Manling Li, The University of Illinois Urbana-Champaign
Website: `https://limanling.github.io/`
Manling Li is a fourth-year Ph.D. student at the Computer Science Department of UIUC. Manling has won the Best Demo Paper Award at ACL'20, the Best Demo Paper Award at NAACL'21, C.L. Dave and Jane W.S. Liu Award, and has been selected as Mavis Future Faculty Fellow. She is a recipient of Microsoft Research PhD Fellowship.

Ben Zhou, University of Pennsylvania
Website: `http://xuanyu.me/`
Ben Zhou is a third-year Ph.D. student at the Department of Computer and Information Science, UPenn. Ben's research interests are distant supervision extraction and experiential knowledge reasoning. He is a recipient of the ENIAC fellowship from the UPenn, and a finalist of the CRA outstanding undergraduate researcher award.

Heng Ji, The University of Illinois Urbana-Champaign & Alexa AI
Website: `https://blender.cs.illinois.edu/hengji.html`
Heng Ji is a Professor at the CS Department of UIUC, and an Amazon Scholar. Her research interests focus on NLP, especially on Multimedia Multilingual IE, Knowledge Base Population and Knowledge-driven Generation. She has received "AI's 10 to Watch" Award, NSF CAREER, Google Research Award, IBM Watson Faculty Award, Bosch Research Award, and Amazon AWS Award, ACL2020 Best Demo Paper Award, and NAACL2021 Best Demo Award.

Dan Roth, University of Pennsylvania & AWS AI Labs
Website: `http://www.cis.upenn.edu/~danroth/`
Dan Roth is the Eduardo D. Glandt Distinguished Professor at the Department of CIS, UPenn, the NLP Lead at AWS AI, and a Fellow of the AAAS, ACM, AAAI, and ACL. In 2017 Roth received the John McCarthy Award. Roth has published broadly in ML, NLP, KRR, and learning theory, and has given keynote talks and tutorials in all ACL and AAAI major conferences. Roth was the Editor-in-Chief of JAIR until 2017, and the program chair of AAAI'11, ACL'03 and CoNLL'02.

T4 - Human-Centered Evaluation of Explanations

Jordan Boyd-Graber, Samuel Carton, Shi Feng, Vera Liao, Tania Lombrozo, Alison Smith-Renner, Chenhao Tan

Introductory

Sunday, July 10, 2022 - 14:00-17:30. Extra Q&A Sessions: 13:30-14:00 and 18:00-18:45 (Ballroom Columbia D)

`https://xai-hcee.github.io/`

The NLP community are increasingly interested in providing explanations for NLP models to help people make sense of model behavior and potentially improve human interaction with models. In addition to computational challenges in generating these explanations, evaluations of the generated explanations require human-centered perspectives and approaches. This tutorial will provide an overview of human-centered evaluations of explanations. First, we will give a brief introduction to the psychological foundation of explanations as well as types of NLP model explanations and their corresponding presentation, to provide the necessary background. We will then present a taxonomy of human-centered evaluation of explanations and dive into depth in the two categories: 1) evaluation with human-subject studies and 2) evaluation based on human-annotated explanations. We will conclude by discussing future directions. We will also adopt a flipped format to maximize the interactive components for the live audience.

Jordan Boyd-Graber, University of Maryland
Website: `http://boydgraber.org/`
Jordan Boyd-Graber is an associate professor at the University of Maryland, with joint appointments between computer science, the iSchool, language science, and the Institute for Advanced Computer Studies. He has been teaching using a flipped classroom approach since 2013. He and his collaborators helped end the use of perplexity for topic models (Chang et al., 2009), first developed interactive topic models (Hu et al., 2011), and improved word-level analysis of topic model explanations (Lund et al., 2019).

Samuel Carton, University of Chicago
Website: `https://shcarton.github.io`
Samuel Carton is a postdoctoral researcher at the University of Chicago. His interests lie in model interpretability and human-AI interaction.

Shi Feng, University of Chicago
Website: `http://www.shifeng.umiacs.io`
Shi Feng is a postdoctoral researcher at the University of Chicago. His research interests include interpretable NLP, adversarial robustness and alignment.

Vera Liao, Microsoft Research Montreal
Website: `http://www.qveraliao.com/`

Vera Liao is a Principal Researcher at Microsoft Research Montreal, where she is part of the FATE (Fairness, Accountability, Transparency, and Ethics) group. She is an HCI researcher by training, with current interest in human-AI interaction and explainable AI.

Tania Lombrozo, Princeton University
Website: `http://cognition.princeton.edu/`
Tania Lombrozo is the Arthur W. Marks '19 Professor of Psychology at Princeton University. She is a leading expert in understanding explanations.

Alison Smith-Renner, Dataminr
Website: `https://alisonmsmith.github.io/`
Alison Smith-Renner is a Senior Research Scientist in human-AI interaction at Dataminr. Her research interests include explainable and interactive natural language processing from a human-centric perspective.

Chenhao Tan, University of Chicago
Website: `https://chenhaot.com/`
Chenhao Tan is an assistant professor of computer science at the University of Chicago, and is also affiliated with the Harris School of Public Policy. His research interests include natural language processing, human-centered AI, and computational social science.

T5 - Tutorial on Multimodal Machine Learning

Louis-Philippe Morency, Paul Pu Liang, Amir Zadeh

Cutting-edge

Sunday, July 10, 2022 - 14:00-17:30. Extra Q&A Sessions: 13:30-14:00 and 18:00-18:45 (Ballroom Columbia C)

https://cmu-multicomp-lab.github.io/mmml-tutorial/naacl2022/

Multimodal machine learning is a vibrant multi-disciplinary research field that addresses some of the original goals of AI via designing computer agents that are able to demonstrate intelligent capabilities such as understanding, reasoning and planning through integrating and modeling multiple communicative modalities, including linguistic, acoustic, and visual messages. With the initial research on audio-visual speech recognition and more recently with language & vision projects such as image and video captioning, visual question answering, and language-guided reinforcement learning, this research field brings some unique challenges for multimodal researchers given the heterogeneity of the data and the contingency often found between modalities.

This tutorial builds upon the annual course on multimodal machine learning taught at Carnegie Mellon University and is a completely revised version of the previous tutorials on multimodal learning at CVPR, ACL, and ICMI conferences. The present tutorial is based on a revamped taxonomy of the core technical challenges present in multimodal machine learning, centered around these six core challenges: representation, alignment, reasoning, induction, generation and quantification. Recent technical achievements will be presented through the lens of this revamped taxonomy of multimodal core challenges, allowing researchers to understand similarities and differences between approaches and new models. The tutorial is also designed to give a perspective on future research directions in multimodal machine learning.

Louis-Philippe Morency, Language Technologies Institute, Carnegie Mellon University
Website: https://www.cs.cmu.edu/~morency/
Louis-Philippe Morency is an Associate Professor at CMU LTI where he leads the Multimodal Communication and Machine Learning Laboratory. He was formerly research faculty at USC and received his Ph.D. degree from MIT CSAIL. His research focuses on building the computational foundations to enable computers with the abilities to analyze, recognize and predict subtle human communicative behaviors during social interactions. He received diverse awards including AI's 10 to Watch by IEEE Intelligent Systems, NetExplo Award in partnership with UNESCO and 10 best paper awards at IEEE and ACM conferences.

Paul Pu Liang, Machine Learning Department, Carnegie Mellon University
Website: https://www.cs.cmu.edu/~pliang/
Paul Liang is a Ph.D. student in Machine Learning at CMU. His research lies in multimodal machine learning with applications in socially intelligent AI. His research is supported by a Facebook PhD Fellowship and a Center for Machine Learning and Health Fellowship, and has been recognized by awards

at the NeurIPS 2019 workshop on federated learning and ICMI 2017. He regularly organizes courses, workshops, and tutorials on multimodal learning.

Amir Zadeh, Language Technologies Institute, Carnegie Mellon University
Website: `https://www.amir-zadeh.com/`
Amir Zadeh is a Postdoctoral Associate at CMU. Prior to that, he received his Ph.D. from Language Technologies Institute, CMU. His work is focused on multimodal learning, especially modeling multimodal language. He is the creator of several resources in this area including CMU-MOSEAS, CMU-MOSEI, and CMU-MOSI datasets. He has organized workshops on multimodal language and published in ACL, EMNLP, NAACL, CVPR, and ICLR.

T6 - Contrastive Data and Learning for Natural Language Processing

Rui Zhang, Yangfeng Ji, Yue Zhang, Rebecca J. Passonneau

Cutting-edge

Sunday, July 10, 2022 - 14:00-17:30. Extra Q&A Sessions: 13:30-14:00 and 18:00-18:45 (Ballroom Columbia A)

`https://contrastive-nlp-tutorial.github.io/`

Current NLP models heavily rely on effective representation learning algorithms. Contrastive learning is one such technique to learn an embedding space such that similar data sample pairs have close representations while dissimilar samples stay far apart from each other. It can be used in supervised or unsupervised settings using different loss functions to produce task-specific or general-purpose representations. While it has originally enabled the success for vision tasks, recent years have seen a growing number of publications in contrastive NLP. This first line of works not only delivers promising performance improvements in various NLP tasks, but also provides desired characteristics such as task-agnostic sentence representation, faithful text generation, data-efficient learning in zero-shot and few-shot settings, interpretability and explainability.

In this tutorial, we aim to provide a gentle introduction to the fundamentals of contrastive learning approaches and the theory behind them. We then survey the benefits and the best practices of contrastive learning for various downstream NLP applications including Text Classification, Question Answering, Summarization, Text Generation, Interpretability and Explainability, Commonsense Knowledge and Reasoning, Vision-and-Language.

Rui Zhang, Penn State University
Website: `https://ryanzhumich.github.io/`
Rui Zhang is an Assistant Professor in the Computer Science and Engineering Department of Penn State University and a co-director of the PSU NLP Lab. He serves as an Area Chair at NAACL 2021, EMNLP 2021, and NLPCC 2021. He has been working on contrastive learning for few-shot named entity recognition and logic-consistency text generation.

Yangfeng Ji, University of Virginia
Website: `http://yangfengji.net/`
Yangfeng Ji is the William Wulf Assistant Professor in the Department of Computer Science at the University of Virginia, where he leads the Natural Language Processing group. His research interests include building machine learning models for text understanding and generation. His work on entity-driven story generation won an Outstanding Paper Award at NAACL 2018. He is a co-author of an EMNLP 2020 tutorial on The Amazing World of Neural Language Generation.

Yue Zhang, Westlake University
Website: `https://frcchang.github.io/`
Yue Zhang is an Associate Professor at Westlake University. His research interests include NLP and its underlying machine learning algorithms and downstream applications. He is the PC co-chair for CCL (2020) and EMNLP (2022), and the area chairs of ACL (2017/18/19/20/21), COLING (2014/18), NAACL (2015/19/21), EMNLP (2015/17/19/20), EACL (2021) and IJCAI (2021). He won the best paper awards of IALP (2017), COLING (2018) and best paper honorable mention of SemEval (2020). He served as one tutorial chair for ACL 2020 and is the author of EMNLP 2018 tutorial on Joint models for NLP.

Rebecca J. Passonneau, Penn State University
Website: `https://sites.psu.edu/becky`
Rebecca J. Passonneau is a Professor in the Computer Science and Engineering Department of Penn State University and a co-director of the PSU NLP Lab. Her area of research is natural language processing, with a focus on semantics and pragmatics. Her work is reported in over 130 journal and refereed conference publications. She won a Best Paper Runner Up at NAACL 2010. She is a tutorial co-chair for NAACL 2018.

8

Main Conference

Main Conference Program (Overview)

Main Conference Program (Overview): Day 1

7:30 *Registration (Level 3 Foyer)*

7:30-9:00 *Breakfast (Level 3 Pre Function & Foyer)*

8:45-9:15 *Welcome Address (Columbia A, C, D)*

9:15-10:15 *Keynote 1 - Batya Friedman (Columbia A, C, D & 302 Beckler)*

10:15-10:45 Morning Break (Regency A)

10:45-12:15 **Session 1**

Language Generation *Columbia A*	**Summarization** *Columbia C*
Information Extraction *Columbia D*	**Efficient Methods in NLP** *Elwha A*
Dialogue *Elwha B*	**In Person Poster Session** *Regency A & B*

12:15-1:15 Lunch break

1:15-2:15 *Industry Panel: Careers in NLP (Columbia A, C, D & 302 Beckler)*

2:15-2:30 Afternoon Break

2:30-4:00 **Session 2**

Interpretability *Columbia A*	**Semantics** *Columbia C*
Language Resources & Evaluation *Columbia D*	**Machine Translation** *Elwha A*
NLP Applications *Elwha B*	**In Person Poster Session** *Regency A & B*

4:00-4:30 Afternoon break

4:30-5:30 *Plenary Best paper Awards & Land Acknowledgement (Columbia A, C, D & 302 Beckler)*

Main Conference Program (Overview): Day 2

7:30	*Registration (Level 3 Foyer)*		
7:30-8:00	*Breakfast (Between Levels 3 & 5Pre Function & Foyer)*		
8:00-9:00	**Session 3**	**Language Generation** *Columbia A*	**Semantics & Sentiment Analysis** *Columbia C*
		Language Resources & Evaluation *Columbia D*	**Efficient Methods in NLP** *Elwha A*
		NLP Applications *Elwha B*	**Industry Oral Session** *Quinault*
		Virtual Poster Q&A Session *702 - Clearwater*	
9:00-9:15	Morning break		
9:15-10:15	*Panel: The Place of Linguistics & Symbolic Structures (Columbia A, C, D & 302 Beckler)*		
10:15-10:45	Morning break		
10:45-12:15	**Session 4**	**Interpretability & Analysis of Models for NLP** *Columbia A*	**Summarization** *Columbia C*
		Information Retrieval *Columbia D*	**Language Grounding to Vision** *Elwha A*
		Dialogue & Interactive Systems *Elwha B*	**In-Person Poster Session** *Regency A & B*
		SRW Panel Discussion for Starting Researchers *Quinault*	
12:15-2:15	Lunch		
2:15-3:45	**Session 5**	**Ethics, Bias, & Fairness** *Columbia A*	**Sentiment Analysis & Stylistic Analysis** *Columbia C*
		Information Extraction *Columbia D*	**Human-Centered NLP** *Elwha A*
		NLP Applications *Elwha B*	**Industry Oral Session** *Quinault*
		SRW In-Person Poster Session *Regency A & B*	
3:45-4:15	Afternoon Break		
4:15-5:45	**Session 6**	**Language Grounding to Vision** *Columbia A*	**Syntax: Tagging, Chunking & Parsing** *Columbia C*
		Multilinguality *Columbia D*	**Machine Learning for NLP** *Elwha A*
		Question Answering *Elwha B*	**SRW Thesis Proposals Session** *Quinault*
		Industry/Demo In-Person Poster Session *Regency A & B*	**Virtual Poster Q & A Session** *Clearwater*

Main Conference Program (Overview): Day 3

7:30	*Registration (Level 3 Foyer)*		
7:30-8:00	*Breakfast (Between Levels 3 & 5Pre Function & Foyer)*		
8:00-9:00	**Session 7**	**Interpretability & Analysis of Models for NLP** *Columbia A*	**Summarization** *Columbia C*
		Infomration Extraction *Columbia D*	**Machine Translation** *Elwha A*
		Dialogue & Interactive Systems *Elwha B*	**Machine Learning for NLP** *Quinault*
		Virtual Poster Q&A Session *702 - Clearwater*	
9:00-9:15	Morning break		
9:15-10:15	**Session 8**	**Interpretability & Analysis of Models for NLP** *Columbia A*	**Comp. Social Science & Cultural Analytics** *Columbia C*
		Machine Learning & Human-Centered NLP *Columbia D*	**Machine Translation** *Elwha A*
		Dialogue & Interactive Systems *Elwha B*	**Machine Learning for NLP** *Quinault*
		Virtual Poster Session *702 - Clearwater*	
10:15-10:45	Morning break		
10:45-12:15	**Session 9**	**Language Generation** *Columbia A*	**Speech & Phonology, Morphology** *Columbia C*
		Information Extraction *Columbia D*	**Language Resources & Evaluation** *Elwha A*
		NLP Applications *Elwha B*	**Findings In-Person Poster Session** *Regency A & B*
12:15-2:15	Lunch		
2:15-3:45	**Session 10**	**Ethics, Bias, Fairness** *Columbia A*	**Semantics** *Columbia C*
		Linguistic Theories, Cogn. Modeling, Discourse *Columbia D*	**Machine Learning** *Elwha A*
		Question Answering *Elwha B*	**Findings In-Person Poster Session** *Regency A & B*
3:45-4:15	Lunch		
4:15-5:15	*Keynote Talk: Manuel Montes-y-Gomez (Columbia A, C, D & 302 Beckler)*		
5:15-5:45	*Closing Session (Columbia A, C, D & 302 Beckler)*		

Main Conference: Monday, July 11, 2022

Session 1 - 10:45-12:15

Language Generation

10:45-12:15 (Columbia A)

Learning to Transfer Prompts for Text Generation
Junyi Li, Tianyi Tang, Jian-Yun Nie, Ji-Rong Wen and Xin Zhao 10:45-11:00 (Columbia A)
Pretrained language models (PLMs) have made remarkable progress in text generation tasks via fine-tuning. While, it is challenging to fine-tune PLMs in a data-scarce situation. Therefore, it is non-trivial to develop a general and lightweight model that can adapt to various text generation tasks based on PLMs. To fulfill this purpose, the recent prompt-based learning offers a potential solution. In this paper, we improve this technique and propose a novel prompt-based method (PTG) for text generation in a transferable setting. First, PTG learns a set of source prompts for various source generation tasks and then transfers these prompts as target prompts to perform target generation tasks. To consider both task- and instance-level information, we design an adaptive attention mechanism to derive the target prompts. For each data instance, PTG learns a specific target prompt by attending to highly relevant source prompts. In extensive experiments, PTG yields competitive or better results than fine-tuning methods. We release our source prompts as an open resource, where users can add or reuse them to improve new text generation tasks for future research. Code and data can be available at `https://github.com/RUCAIBox/Transfer-Prompts-for-Text-Generation`.

Long-term Control for Dialogue Generation: Methods and Evaluation
Ramya Ramakrishnan, Hashan Buddhika Narangodage, Mauro Schilman, Kilian Q Weinberger and Ryan McDonald 11:00-11:15 (Columbia A)
Current approaches for controlling dialogue response generation are primarily focused on high-level attributes like style, sentiment, or topic. In this work, we focus on constrained long-term dialogue generation, which involves more fine-grained control and requires a given set of control words to appear in generated responses. This setting requires a model to not only consider the generation of these control words in the immediate context, but also produce utterances that will encourage the generation of the words at some time in the (possibly distant) future. We define the problem of constrained long-term control for dialogue generation, identify gaps in current methods for evaluation, and propose new metrics that better measure long-term control. We also propose a retrieval-augmented method that improves performance of long-term controlled generation via logit modification techniques. We show through experiments on three task-oriented dialogue datasets that our metrics better assess dialogue control relative to current alternatives and that our method outperforms state-of-the-art constrained generation baselines.

PPL-MCTS: Constrained Textual Generation Through Discriminator-Guided MCTS Decoding
Antoine Chaffin, Vincent Claveau and Ewa Kijak 11:15-11:30 (Columbia A)
Large language models (LM) based on Transformers allow to generate plausible long texts. In this paper, we explore how this generation can be further controlled at decoding time to satisfy certain constraints (e.g. being non-toxic, conveying certain emotions, using a specific writing style, etc.) without fine-tuning the LM. Precisely, we formalize constrained generation as a tree exploration process guided by a discriminator that indicates how well the associated sequence respects the constraint. This approach, in addition to being easier and cheaper to train than fine-tuning the LM, allows to apply the constraint more finely and dynamically. We propose several original methods to search this generation tree, notably the Monte Carlo Tree Search (MCTS) which provides theoretical guarantees on the search efficiency, but also simpler methods based on re-ranking a pool of diverse sequences using the discriminator scores. These methods are evaluated, with automatic and human-based metrics, on two types of constraints and languages: review polarity and emotion control in French and English. We show that discriminator-guided MCTS decoding achieves state-of-the-art results without having to tune the language model, in both tasks and languages. We also demonstrate that other proposed decoding methods based on re-ranking can be really effective when diversity among the generated propositions is encouraged.

RSTGen: Imbuing Fine-Grained Interpretable Control into Long-FormText Generators
Rilwan Akanni Adewoyin, Ritabrata Dutta and Yulan He 11:30-11:45 (Columbia A)
In this paper, we study the task of improving the cohesion and coherence of long-form text generated by language models. To this end, we propose RSTGen, a framework that utilises Rhetorical Structure Theory (RST), a classical language theory, to control the discourse structure, semantics and topics of generated text. Firstly, we demonstrate our model's ability to control structural discourse and semantic features of generated text in open generation evaluation. Then we experiment on the two challenging long-form text tasks of argument generation and story generation. Evaluation using automated metrics and a metric with high correlation to human evaluation, shows that our model performs competitively against existing models, while offering significantly more controls over generated text than alternative methods.

Robust (Controlled) Table-to-Text Generation with Structure-Aware Equivariance Learning
Fei Wang, Zhewei Xu, Pedro Szekely and Muhao Chen 11:45-12:00 (Columbia A)
Controlled table-to-text generation seeks to generate natural language descriptions for highlighted subparts of a table. Previous SOTA systems still employ a sequence-to-sequence generation method, which merely captures the table as a linear structure and is brittle when table layouts change. We seek to go beyond this paradigm by (1) effectively expressing the relations of content pieces in the table, and (2) making our model robust to content-invariant structural transformations. Accordingly, we propose an equivariance learning framework, which encodes tables with a structure-aware self-attention mechanism. This prunes the full self-attention structure into an order-invariant graph attention that captures the connected graph structure of cells belonging to the same row or column, and it differentiates between relevant cells and irrelevant cells from the structural perspective. Our framework also modifies the positional encoding mechanism to preserve the relative position of tokens in the same cell but enforce position invariance among different cells. Our technology is free to be plugged into existing table-to-text generation models, and has improved T5-based models to offer better performance on ToTTo and HiTab. Moreover, on a harder version of ToTTo, we preserve promising performance, while previous SOTA systems, even with transformation-based data augmentation, have seen significant performance drops.

TRUE: Re-evaluating Factual Consistency Evaluation
Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim and Yossi Matias 12:00-12:15 (Columbia A)
Grounded text generation systems often generate text that contains factual inconsistencies, hindering their real-world applicability. Automatic factual consistency evaluation may help alleviate this limitation by accelerating evaluation cycles, filtering inconsistent outputs and augmenting training data. While attracting increasing attention, such evaluation metrics are usually developed and evaluated in silo for a single task or dataset, slowing their adoption. Moreover, previous meta-evaluation protocols focused on system-level correlations with human annotations, which leave the example-level accuracy of such metrics unclear. In this work, we introduce TRUE: a comprehensive survey and assessment of factual consistency metrics on a standardized collection of existing texts from diverse tasks, manually annotated for factual consistency. Our standardization enables an example-level meta-evaluation protocol that is more actionable and interpretable than previously reported correlations, yielding clearer quality measures. Across diverse state-of-the-art metrics and 11 datasets we find that large-scale NLI and question generation-and-answering-based approaches achieve strong and complementary results. We recommend those methods as a starting point for model and metric developers, and hope TRUE will foster progress towards even better evaluation methods.

Summarization

10:45-12:15 (Columbia C)

AnswerSumm: A Manually-Curated Dataset and Pipeline for Answer Summarization
Alexander Fabbri, Xiaojian Wu, Srini Iyer, Haoran Li and Mona T. Diab 10:45-11:00 (Columbia C)
Community Question Answering (CQA) fora such as Stack Overflow and Yahoo! Answers contain a rich resource of answers to a wide range of community-based questions. Each question thread can receive a large number of answers with different perspectives. One goal of answer summarization is to produce a summary that reflects the range of answer perspectives. A major obstacle for this task is the absence of a dataset to provide supervision for producing such summaries. Recent works propose heuristics to create such data, but these are often noisy and do not cover all answer perspectives present. This work introduces a novel dataset of 4,631 CQA threads for answer summarization curated by professional linguists. Our pipeline gathers annotations for all subtasks of answer summarization, including relevant answer sentence selection, grouping these sentences based on perspectives, summarizing each perspective, and producing an overall summary. We analyze and benchmark state-of-the-art models on these subtasks and introduce a novel unsupervised approach for multi-perspective data augmentation that boosts summarization performance according to automatic evaluation. Finally, we propose reinforcement learning rewards to improve factual consistency and answer coverage and analyze areas for improvement.

QAFactEval: Improved QA-Based Factual Consistency Evaluation for Summarization
Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu and Caiming Xiong 11:00-11:15 (Columbia C)
Factual consistency is an essential quality of text summarization models in practical settings. Existing work in evaluating this dimension can be broadly categorized into two lines of research, entailment-based and question answering (QA)-based metrics, and different experimental setups often lead to contrasting conclusions as to which paradigm performs the best. In this work, we conduct an extensive comparison of entailment and QA-based metrics, demonstrating that carefully choosing the components of a QA-based metric, especially question generation and answerability classification, is critical to performance. Building on those insights, we propose an optimized metric, which we call QAFactEval, that leads to a 14% average improvement over previous QA-based metrics on the SummaC factual consistency benchmark, and also outperforms the best-performing entailment-based metric. Moreover, we find that QA-based and entailment-based metrics can offer complementary signals and be combined into a single metric for a further performance boost.

Re-Examining System-Level Correlations of Automatic Summarization Evaluation Metrics
Daniel Deutsch, Rotem Dror and Dan Roth 11:15-11:30 (Columbia C)
How reliably an automatic summarization evaluation metric replicates human judgments of summary quality is quantified by system-level correlations. We identify two ways in which the definition of the system-level correlation is inconsistent with how metrics are used to evaluate systems in practice and propose changes to rectify this disconnect. First, we calculate the system score for an automatic metric using the full test set instead of the subset of summaries judged by humans, which is currently standard practice. We demonstrate how this small change leads to more precise estimates of system-level correlations. Second, we propose to calculate correlations only on pairs of systems that are separated by small differences in automatic scores which are commonly observed in practice. This allows us to demonstrate that our best estimate of the correlation of ROUGE to human judgments is near 0 in realistic scenarios. The results from the analyses point to the need to collect more high-quality human judgments and to improve automatic metrics when differences in system scores are small.

Massive-scale Decoding for Text Generation using Lattices
Jiacheng Xu, Siddhartha Jonnalagadda and Greg Durrett 11:30-11:45 (Columbia C)
Conditional neural text generation models generate high-quality outputs, but often concentrate around a mode when what we really want is a diverse set of options. We present a search algorithm to construct lattices encoding a massive number of generation options. First, we restructure decoding as a best-first search, which explores the space differently than beam search and improves efficiency by avoiding pruning paths. Second, we revisit the idea of hypothesis recombination: we can identify pairs of similar generation candidates during search and merge them as an approximation. On both summarization and machine translation, we show that our algorithm encodes thousands of diverse options that remain grammatical and high-quality into one lattice. This algorithm provides a foundation for building downstream generation applications on top of massive-scale diverse outputs.

FactGraph: Evaluating Factuality in Summarization with Semantic Graph Representations
Leonardo F. R. Ribeiro, Mengwen Liu, Iryna Gurevych, Markus Dreyer and Mohit Bansal 11:45-12:00 (Columbia C)
Despite recent improvements in abstractive summarization, most current approaches generate summaries that are not factually consistent with the source document, severely restricting their trust and usage in real-world applications. Recent works have shown promising improvements in factuality error identification using text or dependency arc entailments; however, they do not consider the entire semantic graph simultaneously. To this end, we propose FactGraph, a method that decomposes the document and the summary into structured meaning representations (MR), which are more suitable for factuality evaluation. MRs describe core semantic concepts and their relations, aggregating the main content in both document and summary in a canonical form, and reducing data sparsity. FactGraph encodes such graphs using a graph encoder augmented with structure-aware adapters to capture interactions among the concepts based on the graph connectivity, along with text representations using an adapter-based text encoder. Experiments on different benchmarks for evaluating factuality show that FactGraph outperforms previous approaches by up to 15%. Furthermore, FactGraph improves performance on identifying content verifiability errors and better captures subsentence-level factual inconsistencies.

CONFIT: Toward Faithful Dialogue Summarization with Linguistically-Informed Contrastive Fine-tuning
Xiangru Tang, Arjun Nair, Borui Wang, Bingyao Wang, Jai Amit Desai, Aaron Wade, Haoran Li, Asli Celikyilmaz, Yashar Mehdad and Dragomir Radev 12:00-12:15 (Columbia C)
Factual inconsistencies in generated summaries severely limit the practical applications of abstractive dialogue summarization. Although significant progress has been achieved by using pre-trained neural language models, substantial amounts of hallucinated content are found during the human evaluation. In this work, we first devised a typology of factual errors to better understand the types of hallucinations generated by current models and conducted human evaluation on popular dialog summarization dataset. We further propose a training strategy that improves the factual consistency and overall quality of summaries via a novel contrastive fine-tuning, called CONFIT. To tackle top factual errors from our annotation, we introduce additional contrastive loss with carefully designed hard negative samples and self-supervised dialogue-specific loss to capture the key information between speakers. We show that our model significantly reduces all kinds of factual errors on both SAMSum dialogue summarization and AMI meeting summarization. On both datasets, we achieve significant improvements over state-of-the-art baselines using both automatic metrics, ROUGE and BARTScore, and human evaluation.

Information Extraction

10:45-12:15 (Columbia D)

An Enhanced Span-based Decomposition Method for Few-Shot Sequence Labeling
Peiyi Wang, Runxin Xu, Tianyu Liu, Qingyu Zhou, Yunbo Cao, Baobao Chang and Zhifang Sui 10:45-11:00 (Columbia D)
Few-Shot Sequence Labeling (FSSL) is a canonical paradigm for the tagging models, e.g., named entity recognition and slot filling, to generalize on an emerging, resource-scarce domain. Recently, the metric-based meta-learning framework has been recognized as a promising approach for FSSL. However, most prior works assign a label to each token based on the token-level similarities, which ignores the integrality of named entities or slots. To this end, in this paper, we propose ESD, an Enhanced Span-based Decomposition method for FSSL. ESD formulates FSSL as a span-level matching problem between test query and supporting instances. Specifically, ESD decomposes the span matching problem into a series of span-level procedures, mainly including enhanced span representation, class prototype aggregation and span conflicts resolution. Extensive experiments show that ESD achieves the new state-of-the-art results on two popular FSSL benchmarks, FewNERD and SNIPS, and is proven to be more robust in the noisy and nested tagging scenarios.

Modeling Multi-Granularity Hierarchical Features for Relation Extraction
Xinnian Liang, Shuangzhi Wu, Mu Li and Zhoujun Li 11:00-11:15 (Columbia D)
Relation extraction is a key task in Natural Language Processing (NLP), which aims to extract relations between entity pairs from given texts. Recently, relation extraction (RE) has achieved remarkable progress with the development of deep neural networks. Most existing research focuses on constructing explicit structured features using external knowledge such as knowledge graph and dependency tree. In this paper, we propose a novel method to extract multi-granularity features based solely on the original input sentences. We show that effective structured features can be attained even without external knowledge. Three kinds of features based on the input sentences are fully exploited, which are in entity mention level, segment level, and sentence level. All the three are jointly and hierarchically modeled. We evaluate our method on three public benchmarks: SemEval 2010 Task 8, Tacred, and Tacred Revisited. To verify the effectiveness, we apply our method to different encoders such as LSTM and BERT. Experimental results show that our method significantly outperforms existing state-of-the-art models that even use external knowledge. Extensive analyses demonstrate that the performance of our model is contributed by the capture of multi-granularity features and the model of their hierarchical structure.

Contrastive Representation Learning for Cross-Document Coreference Resolution of Events and Entities
Benjamin Hsu and Graham Horwood 11:15-11:30 (Columbia D)
Identifying related entities and events within and across documents is fundamental to natural language understanding. We present an approach to entity and event coreference resolution utilizing contrastive representation learning. Earlier state-of-the-art methods have formulated this problem as a binary classification problem and leveraged large transformers in a cross-encoder architecture to achieve their results. For large collections of documents and corresponding set of n mentions, the necessity of performing n^2 transformer computations in these earlier approaches can be computationally intensive. We show that it is possible to reduce this burden by applying contrastive learning techniques that only require n transformer computations at inference time. Our method achieves state-of-the-art results on a number of key metrics on the ECB+ corpus and is competitive on others.

Cross-Lingual Event Detection via Optimized Adversarial Training
Luis Fernando Guzman-Nateras, Minh Van Nguyen and Thien Huu Nguyen 11:30-11:45 (Columbia D)
In this work, we focus on Cross-Lingual Event Detection where a model is trained on data from a *source* language but its performance is evaluated on data from a second, *target*, language. Most recent works in this area have harnessed the language-invariant qualities displayed by pre-trained Multi-lingual Language Models. Their performance, however, reveals there is room for improvement as the cross-lingual setting entails particular challenges. We employ Adversarial Language Adaptation to train a Language Discriminator to discern between the source and target languages using unlabeled data. The discriminator is trained in an adversarial manner so that the encoder learns to produce refined, language-invariant representations that lead to improved performance. More importantly, we optimize the adversarial training process by only presenting the discriminator with the most informative samples. We base our intuition about what makes a sample informative on two disparate metrics: sample similarity and event presence. Thus, we propose leveraging Optimal Transport as a solution to naturally combine these two distinct information sources into the selection process. Extensive experiments on 8 different language pairs, using 4 languages from unrelated families, show the flexibility and effectiveness of our model that achieves state-of-the-art results.

Learning to Borrow– Relation Representation for Without-Mention Entity-Pairs for Knowledge Graph Completion
Huda Hakami, Mona Hakami, Angrosh Mandya and Danushka Bollegala 11:45-12:00 (Columbia D)
Prior work on integrating text corpora with knowledge graphs (KGs) to improve Knowledge Graph Embedding (KGE) have obtained good performance for entities that co-occur in sentences in text corpora. Such sentences (textual mentions of entity-pairs) are represented as Lexicalised Dependency Paths (LDPs) between two entities. However, it is not possible to represent relations between entities that do not co-occur in a single sentence using LDPs. In this paper, we propose and evaluate several methods to address this problem, where we *borrow* LDPs from the entity pairs that co-occur in sentences in the corpus (i.e. *with mentions* entity pairs) to represent entity pairs that do *not* co-occur in any sentence in the corpus (i.e. *without mention* entity pairs). We propose a supervised borrowing method, *SuperBorrow*, that learns to score the suitability of an LDP to represent a without-mentions entity pair using pre-trained entity embeddings and contextualised LDP representations. Experimental results show that SuperBorrow improves the link prediction performance of multiple widely-used prior KGE methods such as TransE, DistMult, ComplEx and RotatE.

[TACL] **Ultra-fine Entity Typing with Indirect Supervision from Natural Language Inference**
Bangzheng Li, Wenpeng Yin and Muhao Chen 12:00-12:15 (Columbia D)
The task of ultra-fine entity typing (UFET) seeks to predict diverse and free-form words or phrases that describe the appropriate types of entities mentioned in sentences. A key challenge for this task lies in the large amount of types and the scarcity of annotated data per type. Existing systems formulate the task as a multi-way classification problem and train directly or distantly supervised classifiers. This causes two issues: (i) the classifiers do not capture the type semantics since types are often converted into indices; (ii) systems developed in this way are limited to predicting within a pre-defined type set, and often fall short of generalizing to types that are rarely seen or unseen in training. This work presents LITE, a new approach that formulates entity typing as a natural language inference (NLI) problem, making use of (i) the indirect supervision from NLI to infer type information meaningfully represented as textual hypotheses and alleviate the data scarcity issue, as well as (ii) a learning-to-rank objective to avoid the pre-defining of a type set. Experiments show that, with limited training data, LITE obtains state-of-the-art performance on the UFET task. In addition, LITE demonstrates its strong generalizability, by not only yielding best results on other fine-grained entity typing benchmarks, more importantly, a pre-trained LITE system works well on new data containing unseen types.

Efficient Methods in NLP 1

10:45-12:15 (Elwha A)

KroneckerBERT: Significant Compression of Pre-trained Language Models Through Kronecker Decomposition and Knowledge Distillation
Marzieh S. Tahaei, Ella Charlaix, Vahid Partovi Nia, Ali Ghodsi and Mehdi Rezagholizadeh 10:45-11:00 (Elwha A)
The development of over-parameterized pre-trained language models has made a significant contribution toward the success of natural language processing. While over-parameterization of these models is the key to their generalization power, it makes them unsuitable for deployment on low-capacity devices. We push the limits of state-of-the-art Transformer-based pre-trained language model compression using Kronecker decomposition. We present our KroneckerBERT, a compressed version of the BERT_BASE model obtained by compressing the embedding layer and the linear mappings in the multi-head attention, and the feed-forward network modules in the Transformer layers. Our KroneckerBERT is trained via a very efficient two-stage knowledge distillation scheme using far fewer data samples than state-of-the-art models like MobileBERT and TinyBERT. We evaluate the performance of KroneckerBERT on well-known NLP benchmarks. We show that our KroneckerBERT with compression factors of 7.7x and 21x outperforms state-of-the-art compression methods on the GLUE and SQuAD benchmarks. In particular, using only 13% of the teacher model parameters, it retain more than 99% of the accuracy on the majority of GLUE tasks.

Meta Learning for Natural Language Processing: A Survey
Hung-yi Lee, Shang-Wen Li and Thang Vu 11:00-11:15 (Elwha A)
Deep learning has been the mainstream technique in the natural language processing (NLP) area. However, deep learning requires many labeled data and is less generalizable across domains. Meta-learning is an arising field in machine learning. It studies approaches to learning better learning algorithms and aims to improve algorithms in various aspects, including data efficiency and generalizability. The efficacy of meta-learning has been shown in many NLP tasks, but there is no systematic survey of these approaches in NLP, which hinders more researchers from joining the field. Our goal with this survey paper is to offer researchers pointers to relevant meta-learning works in NLP and attract more attention from the NLP community to drive future innovation. This paper first introduces the general concepts of meta-learning and the common approaches. Then we summarize task construction settings, applications of meta-learning for various NLP problems and review the development of meta-learning in the NLP community.

On Transferability of Prompt Tuning for Natural Language Processing
Yusheng Su, Xiaozhi Wang, Yujia Qin, Chi-Min Chan, Yankai Lin, Huadong Wang, Kaiyue Wen, Zhiyuan Liu, Peng Li, Juanzi Li, Lei Hou, Maosong Sun and Jie Zhou 11:15-11:30 (Elwha A)
Prompt tuning (PT) is a promising parameter-efficient method to utilize extremely large pre-trained language models (PLMs), which can achieve comparable performance to full-parameter fine-tuning by only tuning a few soft prompts. However, PT requires much more training time than fine-tuning. Intuitively, knowledge transfer can help to improve the efficiency. To explore whether we can improve PT via prompt transfer, we empirically investigate the transferability of soft prompts across different downstream tasks and PLMs in this work. We find that (1) in zero-shot setting, trained soft prompts can effectively transfer to similar tasks on the same PLM and also to other PLMs with a cross-model projector trained on similar tasks; (2) when used as initialization, trained soft prompts of similar tasks and projected prompts of other PLMs can significantly accelerate training and also improve the performance of PT. Moreover, to explore what decides prompt transferability, we investigate various transferability indicators and find that the overlapping rate of activated neurons strongly reflects the transferability, which suggests how the prompts stimulate PLMs is essential. Our findings show that prompt transfer is promising for improving PT, and further research shall focus more on prompts' stimulation to PLMs. The source code can be obtained from https://github.com/thunlp/Prompt-Transferability.

Sparse Distillation: Speeding Up Text Classification by Using Bigger Student Models
Qinyuan Ye, Madian Khabsa, Mike Lewis, Sinong Wang, Xiang Ren and Aaron Jaech 11:30-11:45 (Elwha A)
Distilling state-of-the-art transformer models into lightweight student models is an effective way to reduce computation cost at inference time. The student models are typically compact transformers with fewer parameters, while expensive operations such as self-attention persist. Therefore, the improved inference speed may still be unsatisfactory for real-time or high-volume use cases. In this paper, we aim to further push the limit of inference speed by distilling teacher models into bigger, sparser student models – bigger in that they scale up to billions of parameters; sparser in that most of the model parameters are n-gram embeddings. Our experiments on six single-sentence text classification tasks show that these student models retain 97% of the RoBERTa-Large teacher performance on average, and meanwhile achieve up to 600x speed-up on both GPUs and CPUs at inference time. Further investigation reveals that our pipeline is also helpful for sentence-pair classification tasks, and in domain generalization settings.

Sketching as a Tool for Understanding and Accelerating Self-attention for Long Sequences
Yifan Chen, Qi Zeng, Dilek Hakkani-Tur, Di Jin, Heng Ji and Yun Yang 11:45-12:00 (Elwha A)
Transformer-based models are not efficient in processing long sequences due to the quadratic space and time complexity of the self-attention modules. To address this limitation, Linformer and Informer reduce the quadratic complexity to linear (modulo logarithmic factors) via low-dimensional projection and row selection, respectively. These two models are intrinsically connected, and to understand their connection we introduce a theoretical framework of matrix sketching. Based on the theoretical analysis, we propose Skeinformer to accelerate self-attention and further improve the accuracy of matrix approximation to self-attention with column sampling, adaptive row normalization and pilot sampling reutilization. Experiments on the Long Range Arena benchmark demonstrate that our methods outperform alternatives with a

consistently smaller time/space footprint.

FNet: Mixing Tokens with Fourier Transforms
James Lee-Thorp, Joshua Ainslie, Ilya Eckstein and Santiago Ontanon 12:00-12:15 (Elwha A)
We show that Transformer encoder architectures can be sped up, with limited accuracy costs, by replacing the self-attention sublayers with simple linear transformations that "mix" input tokens. Most surprisingly, we find that replacing the self-attention sublayer in a Transformer encoder with a standard, unparameterized Fourier Transform achieves 92-97% of the accuracy of BERT counterparts on the GLUE benchmark, but trains 80% faster on GPUs and 70% faster on TPUs at standard 512 input lengths. At longer input lengths, our FNet model is significantly faster: when compared to the "efficient Transformers" on the Long Range Arena benchmark, FNet matches the accuracy of the most accurate models, while outpacing the fastest models across all sequence lengths on GPUs (and across relatively shorter lengths on TPUs). Finally, FNet has a light memory footprint and is particularly efficient at smaller model sizes; for a fixed speed and accuracy budget, small FNet models outperform Transformer counterparts.

Dialogue

10:45-12:15 (Elwha B)

LUNA: Learning Slot-Turn Alignment for Dialogue State Tracking
Yifan Wang, Jing Zhao, Junwei Bao, Chaoqun Duan, Youzheng Wu and Xiaodong He 10:45-11:00 (Elwha B)
Dialogue state tracking (DST) aims to predict the current dialogue state given the dialogue history. Existing methods generally exploit the utterances of all dialogue turns to assign value for each slot. This could lead to suboptimal results due to the information introduced from irrelevant utterances in the dialogue history, which may be useless and can even cause confusion. To address this problem, we propose LUNA, a SLot-TUrN Alignment enhanced approach. It first explicitly aligns each slot with its most relevant utterance, then further predicts the corresponding value based on this aligned utterance instead of all dialogue utterances. Furthermore, we design a slot ranking auxiliary task to learn the temporal correlation among slots which could facilitate the alignment. Comprehensive experiments are conducted on three multi-domain task-oriented dialogue datasets, MultiWOZ 2.0, MultiWOZ 2.1, and MultiWOZ 2.2. The results show that LUNA achieves new state-of-the-art results on these datasets.

Stylized Knowledge-Grounded Dialogue Generation via Disentangled Template Rewriting
Qingfeng Sun, Can Xu, Huang Hu, Yujing Wang, Jian Miao, Xiubo Geng, Yining Chen, Fei Xu and Daxin Jiang 11:00-11:15 (Elwha B)
Current Knowledge-Grounded Dialogue Generation (KDG) models specialize in producing rational and factual responses. However, to establish long-term relationships with users, the KDG model needs the capability to generate responses in a desired style or attribute. Thus, we study a new problem: Stylized Knowledge-Grounded Dialogue Generation (SKDG). It presents two challenges: (1) How to train a SKDG model where no <context, knowledge, stylized response> triples are available. (2) How to cohere with context and preserve the knowledge when generating a stylized response. In this paper, we propose a novel disentangled template rewriting (DTR) method which generates responses via combing disentangled style templates (from monolingual stylized corpus) and content templates (from KDG corpus). The entire framework is end-to-end differentiable and learned without supervision. Extensive experiments on two benchmarks indicate that DTR achieves a significant improvement on all evaluation metrics compared with previous state-of-the-art stylized dialogue generation methods. Besides, DTR achieves comparable performance with the state-of-the-art KDG methods in standard KDG evaluation setting.

Enhance Incomplete Utterance Restoration by Joint Learning Token Extraction and Text Generation
Shumpei Inoue, Tsungwei Liu, Son Hong Nguyen and Minh-Tien Nguyen 11:15-11:30 (Elwha B)
This paper introduces a model for incomplete utterance restoration (IUR) called JET (Joint learning token Extraction and Text generation). Different from prior studies that only work on extraction or abstraction datasets, we design a simple but effective model, working for both scenarios of IUR. Our design simulates the nature of IUR, where omitted tokens from the context contribute to restoration. From this, we construct a Picker that identifies the omitted tokens. To support the picker, we design two label creation methods (soft and hard labels), which can work in cases of no annotation data for the omitted tokens. The restoration is done by using a Generator with the help of the Picker on joint learning. Promising results on four benchmark datasets in extraction and abstraction scenarios show that our model is better than the pretrained T5 and non-generative language model methods in both rich and limited training data settings.

Multi2WOZ: A Robust Multilingual Dataset and Conversational Pretraining for Task-Oriented Dialog
Chia-Chien Hung, Anne Lauscher, Ivan Vulić, Simone Paolo Ponzetto and Goran Glavaš 11:30-11:45 (Elwha B)
Research on (multi-domain) task-oriented dialog (TOD) has predominantly focused on the English language, primarily due to the shortage of robust TOD datasets in other languages, preventing the systematic investigation of cross-lingual transfer for this crucial NLP application area. In this work, we introduce Multi2WOZ, a new multilingual multi-domain TOD dataset, derived from the well-established English dataset MultiWOZ, that spans four typologically diverse languages: Chinese, German, Arabic, and Russian. In contrast to concurrent efforts, Multi2WOZ contains gold-standard dialogs in target languages that are directly comparable with development and test portions of the English dataset, enabling reliable and comparative estimates of cross-lingual transfer performance for TOD. We then introduce a new framework for multilingual conversational specialization of pretrained language models (PrLMs) that aims to facilitate cross-lingual transfer for arbitrary downstream TOD tasks. Using such conversational PrLMs specialized for concrete target languages, we systematically benchmark a number of zero-shot and few-shot cross-lingual transfer approaches on two standard TOD tasks: Dialog State Tracking and Response Retrieval. Our experiments show that, in most setups, the best performance entails the combination of (i) conversational specialization in the target language and (ii) few-shot transfer for the concrete TOD task. Most importantly, we show that our conversational specialization in the target language allows for an exceptionally sample-efficient few-shot transfer for downstream TOD tasks.

[TACL] **Reducing conversational agents' overconfidence through linguistic calibration**
Mielke, Arthur Szlam and Emily Dinan,Y-Lan Boureau 11:45-12:00 (Elwha B)
While improving neural dialogue agents' factual accuracy is the object of much research, another important aspect of communication, less studied in the setting of neural dialogue, is transparency about ignorance. In this work, we analyze to what extent state-of-the-art chit-chat models are linguistically calibrated in the sense that their verbalized expression of doubt (or confidence) matches the likelihood that the model's responses are factually incorrect (or correct). We find that these models are poorly calibrated, yet we show that the representations within the models can be used to accurately predict likelihood of correctness. By incorporating these metacognitive features into the training of a controllable generation model, we obtain a dialogue agent with greatly improved linguistic calibration.

Intent Detection and Discovery from User Logs via Deep Semi-Supervised Contrastive Clustering
Rajat Kumar, Mayur Patidar, Vaibhav Varshney, Lovekesh Vig and Gautam Shroff 12:00-12:15 (Elwha B)

Intent Detection is a crucial component of Dialogue Systems wherein the objective is to classify a user utterance into one of multiple pre-defined intents. A pre-requisite for developing an effective intent identifier is a training dataset labeled with all possible user intents. However, even skilled domain experts are often unable to foresee all possible user intents at design time and for practical applications, novel intents may have to be inferred incrementally on-the-fly from user utterances. Therefore, for any real-world dialogue system, the number of intents increases over time and new intents have to be discovered by analyzing the utterances outside the existing set of intents. In this paper, our objective is to i) detect known intent utterances from a large number of unlabeled utterance samples given a few labeled samples and ii) discover new unknown intents from the remaining unlabeled samples. Existing SOTA approaches address this problem via alternate representation learning and clustering wherein pseudo labels are used for updating the representations and clustering is used for generating the pseudo labels. Unlike existing approaches that rely on epoch wise cluster alignment, we propose an end-to-end deep contrastive clustering algorithm that jointly updates model parameters and cluster centers via supervised and self-supervised learning and optimally utilizes both labeled and unlabeled data. Our proposed approach outperforms competitive baselines on five public datasets for both settings: (i) where the number of undiscovered intents are known in advance, and (ii) where the number of intents are estimated by an algorithm. We also propose a human-in-the-loop variant of our approach for practical deployment which does not require an estimate of new intents and outperforms the end-to-end approach.

In-person Poster Session 1

10:45-12:15 (Regency A & B)

Political Ideology and Polarization: A Multi-dimensional Approach
Barea Sinno, Bernardo Oviedo, Katherine Atwell, Malihe Alikhani and Junyi Jessy Li 10:45-12:15 (Regency A & B)
Analyzing ideology and polarization is of critical importance in advancing our grasp of modern politics. Recent research has made great strides towards understanding the ideological bias (i.e., stance) of news media along the left-right spectrum. In this work, we instead take a novel and more nuanced approach for the study of ideology based on its left or right positions on the issue being discussed. Aligned with the theoretical accounts in political science, we treat ideology as a multi-dimensional construct, and introduce the first diachronic dataset of news articles whose ideological positions are annotated by trained political scientists and linguists at the paragraph level. We showcase that, by controlling for the author's stance, our method allows for the quantitative and temporal measurement and analysis of polarization as a multidimensional ideological distance. We further present baseline models for ideology prediction, outlining a challenging task distinct from stance detection.

Counterfactually Augmented Data and Unintended Bias: The Case of Sexism and Hate Speech Detection
Indira Sen, Mattia Samory, Claudia Wagner and Isabelle Augenstein 10:45-12:15 (Regency A & B)
Counterfactually Augmented Data (CAD) aims to improve out-of-domain generalizability, an indicator of model robustness. The improvement is credited to promoting core features of the construct over spurious artifacts that happen to correlate with it. Yet, over-relying on core features may lead to unintended model bias. Especially, construct-driven CAD—perturbations of core features—may induce models to ignore the context in which core features are used. Here, we test models for sexism and hate speech detection on challenging data: non-hate and non-sexist usage of identity and gendered terms. On these hard cases, models trained on CAD, especially construct-driven CAD, show higher false positive rates than models trained on the original, unperturbed data. Using a diverse set of CAD—construct-driven and construct-agnostic—reduces such unintended bias.

Combining Humor and Sarcasm for Improving Political Parody Detection
Xiao Ao, Danae Sanchez Villegas, Daniel Preotiuc-Pietro and Nikolaos Aletras 10:45-12:15 (Regency A & B)
Parody is a figurative device used for mimicking entities for comedic or critical purposes. Parody is intentionally humorous and often involves sarcasm. This paper explores jointly modelling these figurative tropes with the goal of improving performance of political parody detection in tweets. To this end, we present a multi-encoder model that combines three parallel encoders to enrich parody-specific representations with humor and sarcasm information. Experiments on a publicly available data set of political parody tweets demonstrate that our approach outperforms previous state-of-the-art methods.

Measuring Fairness with Biased Rulers: A Comparative Study on Bias Metrics for Pre-trained Language Models
Pieter Delobelle, Ewoenam Kwaku Tokpo, Toon Calders and Bettina Berendt 10:45-12:15 (Regency A & B)
An increasing awareness of biased patterns in natural language processing resources such as BERT has motivated many metrics to quantify 'bias' and 'fairness' in these resources. However, comparing the results of different metrics and the works that evaluate with such metrics remains difficult, if not outright impossible. We survey the literature on fairness metrics for pre-trained language models and experimentally evaluate compatibility, including both biases in language models and in their downstream tasks. We do this by combining traditional literature survey, correlation analysis and empirical evaluations. We find that many metrics are not compatible with each other and highly depend on (i) templates, (ii) attribute and target seeds and (iii) the choice of embeddings. We also see no tangible evidence of intrinsic bias relating to extrinsic bias. These results indicate that fairness or bias evaluation remains challenging for contextualized language models, among other reasons because these choices remain subjective. To improve future comparisons and fairness evaluations, we recommend to avoid embedding-based metrics and focus on fairness evaluations in downstream tasks.

Recognition of They/Them as Singular Personal Pronouns in Coreference Resolution
Connor Baumler and Rachel Rudinger 10:45-12:15 (Regency A & B)
As using they/them as personal pronouns becomes increasingly common in English, it is important that coreference resolution systems work as well for individuals who use personal "they" as they do for those who use gendered personal pronouns. We introduce a new benchmark for coreference resolution systems which evaluates singular personal "they" recognition. Using these WinoNB schemas, we evaluate a number of publicly available coreference resolution systems and confirm their bias toward resolving "they" pronouns as plural.

Using Natural Sentence Prompts for Understanding Biases in Language Models
Sarah Alnegheimish, Alicia Guo and Yi Sun 10:45-12:15 (Regency A & B)
Evaluation of biases in language models is often limited to synthetically generated datasets. This dependence traces back to the need of prompt-style dataset to trigger specific behaviors of language models. In this paper, we address this gap by creating a prompt dataset with respect to occupations collected from real-world natural sentences present in Wikipedia. We aim to understand the differences between using template-based prompts and natural sentence prompts when studying gender-occupation biases in language models. We find bias evaluations are very sensitive to the design choices of template prompts, and we propose using natural sentence prompts as a way of more systematically using real-world sentences to move away from design decisions that may bias the results.

Towards Process-Oriented, Modular, and Versatile Question Generation that Meets Educational Needs
Xu Wang, Simin Fan, Jessica Houghton and Lu Wang 10:45-12:15 (Regency A & B)
NLP-powered automatic question generation (QG) techniques carry great pedagogical potential of saving educators' time and benefiting student learning. Yet, QG systems have not been widely adopted in classrooms to date. In this work, we aim to pinpoint key impediments and investigate how to improve the usability of automatic QG techniques for educational purposes by understanding how instructors construct questions and identifying touch points to enhance the underlying NLP models. We perform an in-depth need finding study with 11 instructors across 7 different universities, and summarize their thought processes and needs when creating questions. While instructors show great interests in using NLP systems to support question design, none of them has used such tools in practice. They resort to multiple sources of information, ranging from domain knowledge to students' misconceptions, all of which missing from today's QG systems. We argue that building effective human-NLP collaborative QG systems that emphasize instructor control and explainability is imperative for real-world adoption. We call for QG systems to provide process-oriented support, use modular design, and handle diverse sources of input.

Machine-in-the-Loop Rewriting for Creative Image Captioning
Vishakh Padmakumar and He He 10:45-12:15 (Regency A & B)
Machine-in-the-loop writing aims to build models that assist humans to accomplish their writing tasks more effectively. Prior work has found that providing users a machine-written draft or sentence-level continuations has limited success since the generated text tends to deviate from users' intention. To allow the user to retain control over the content, we train a rewriting model that, when prompted, modifies specified spans of text within the user's original draft to introduce descriptive and figurative elements in the text. We evaluate the model on its ability to collaborate with humans on the task of creative image captioning. On a user study through Amazon Mechanical Turk, our model is rated to be more helpful by users than a baseline infilling language model. In addition, third-party evaluation shows that users write more descriptive and figurative captions when collaborating with our model compared to completing the task alone. However, the improvement is not uniform across user groups: the model is more helpful to skilled users, which risks widening the gap between skilled and novice users, highlighting a need for careful, user-centric evaluation of interactive systems.

What Makes a Good and Useful Summary? Incorporating Users in Automatic Summarization Research
Maartje Ter Hoeve, Julia Kiseleva and Maarten de Rijke 10:45-12:15 (Regency A & B)
Automatic text summarization has enjoyed great progress over the years and is used in numerous applications, impacting the lives of many. Despite this development, there is little research that meaningfully investigates how the current research focus in automatic summarization aligns with users' needs. To bridge this gap, we propose a survey methodology that can be used to investigate the needs of users of automatically generated summaries. Importantly, these needs are dependent on the target group. Hence, we design our survey in such a way that it can be easily adjusted to investigate different user groups. In this work we focus on university students, who make extensive use of summaries during their studies. We find that the current research directions of the automatic summarization community do not fully align with students' needs. Motivated by our findings, we present ways to mitigate this mismatch in future research on automatic summarization: we propose research directions that impact the design, the development and the evaluation of automatically generated summaries.

Towards Process-Oriented, Modular, and Versatile Question Generation that Meets Educational Needs
Xu Wang, Simin Fan, Jessica Houghton and Lu Wang 10:45-12:15 (Regency A & B)
NLP-powered automatic question generation (QG) techniques carry great pedagogical potential of saving educators' time and benefiting student learning. Yet, QG systems have not been widely adopted in classrooms to date. In this work, we aim to pinpoint key impediments and investigate how to improve the usability of automatic QG techniques for educational purposes by understanding how instructors construct questions and identifying touch points to enhance the underlying NLP models. We perform an in-depth need finding study with 11 instructors across 7 different universities, and summarize their thought processes and needs when creating questions. While instructors show great interests in using NLP systems to support question design, none of them has used such tools in practice. They resort to multiple sources of information, ranging from domain knowledge to students' misconceptions, all of which missing from today's QG systems. We argue that building effective human-NLP collaborative QG systems that emphasize instructor control and explainability is imperative for real-world adoption. We call for QG systems to provide process-oriented support, use modular design, and handle diverse sources of input.

Learning Cross-Lingual IR from an English Retriever
Yulong Li, Martin Franz, Md Arafat Sultan, Bhavani Iyer, Young-Suk Lee and Avirup Sil 10:45-12:15 (Regency A & B)
We present DR.DECR (Dense Retrieval with Distillation-Enhanced Cross-Lingual Representation), a new cross-lingual information retrieval (CLIR) system trained using multi-stage knowledge distillation (KD). The teacher of DR.DECR relies on a highly effective but computationally expensive two-stage inference process consisting of query translation and monolingual IR, while the student, DR.DECR, executes a single CLIR step. We teach DR.DECR powerful multilingual representations as well as CLIR by optimizing two corresponding KD objectives. Learning useful representations of non-English text from an English-only retriever is accomplished through a cross-lingual token alignment algorithm that relies on the representation capabilities of the underlying multilingual encoders. In both in-domain and zero-shot out-of-domain evaluation, DR.DECR demonstrates far superior accuracy over direct fine-tuning with labeled CLIR data. It is also the best single-model retriever on the XOR-TyDi benchmark at the time of this writing.

Collective Relevance Labeling for Passage Retrieval
Jihyuk Kim, Minsoo Kim and Seung-won Hwang 10:45-12:15 (Regency A & B)
Deep learning for Information Retrieval (IR) requires a large amount of high-quality query-document relevance labels, but such labels are inherently sparse. Label smoothing redistributes some observed probability mass over unobserved instances, often uniformly, uninformed of the true distribution. In contrast, we propose knowledge distillation for informed labeling, without incurring high computation overheads at evaluation time. Our contribution is designing a simple but efficient teacher model which utilizes collective knowledge, to outperform state-of-the-arts distilled from a more complex teacher model. Specifically, we train up to ×8 faster than the state-of-the-art teacher, while distilling the rankings better. Our code is publicly available at `https://github.com/jihyukkim-nlp/CollectiveKD`.

Improving Neural Models for Radiology Report Retrieval with Lexicon-based Automated Annotation
Luyao Shi, Tanveer Syeda-mahmood and Tyler Baldwin 10:45-12:15 (Regency A & B)
Many clinical informatics tasks that are based on electronic health records (EHR) need relevant patient cohorts to be selected based on findings, symptoms and diseases. Frequently, these conditions are described in radiology reports which can be retrieved using information retrieval (IR) methods. The latest of these techniques utilize neural IR models such as BERT trained on clinical text. However, these methods still lack semantic understanding of the underlying clinical conditions as well as ruled out findings, resulting in poor precision during retrieval. In this paper we combine clinical finding detection with supervised query match learning. Specifically, we use lexicon-driven concept detection to detect relevant findings in sentences. These findings are used as queries to train a Sentence-BERT (SBERT) model using triplet loss on matched and unmatched query-sentence pairs. We show that the proposed supervised training task remarkably improves the retrieval performance of SBERT. The trained model generalizes well to unseen queries and reports from different collections.

Residue-Based Natural Language Adversarial Attack Detection

Vyas Raina and Mark Gales 10:45-12:15 (Regency A & B)
Deep learning based systems are susceptible to adversarial attacks, where a small, imperceptible change at the input alters the model prediction. However, to date the majority of the approaches to detect these attacks have been designed for image processing systems. Many popular image adversarial detection approaches are able to identify adversarial examples from embedding feature spaces, whilst in the NLP domain existing state of the art detection approaches solely focus on input text features, without consideration of model embedding spaces. This work examines what differences result when porting these image designed strategies to Natural Language Processing (NLP) tasks - these detectors are found to not port over well. This is expected as NLP systems have a very different form of input: discrete and sequential in nature, rather than the continuous and fixed size inputs for images. As an equivalent model-focused NLP detection approach, this work proposes a simple sentence-embedding "residue" based detector to identify adversarial examples. On many tasks, it out-performs ported image domain detectors and recent state of the art NLP specific detectors.

Locally Aggregated Feature Attribution on Natural Language Model Understanding
Sheng Zhang, Jin Wang, Haitao Jiang and Rui Song 10:45-12:15 (Regency A & B)
With the growing popularity of deep-learning models, model understanding becomes more important. Much effort has been devoted to demystify deep neural networks for better explainability. Some feature attribution methods have shown promising results in computer vision, especially the gradient-based methods where effectively smoothing the gradients with reference data is the key to a robust and faithful result. However, direct application of these gradient-based methods to NLP tasks is not trivial due to the fact that the input consists of discrete tokens and the “reference” tokens are not explicitly defined. In this work, we propose Locally Aggregated Feature Attribution (LAFA), a novel gradient-based feature attribution method for NLP models. Instead of relying on obscure reference tokens, it smooths gradients by aggregating similar reference texts derived from language model embeddings. For evaluation purpose, we also design experiments on different NLP tasks including Entity Recognition and Sentiment Analysis on public datasets and key words detection on constructed Amazon catalogue dataset. The superior performance of the proposed method is demonstrated through experiments.

Simple Local Attentions Remain Competitive for Long-Context Tasks
*Wenhan Xiong, Barlas Oguz, Anchit Gupta, Xilun Chen, Diana Liskovich, Omer Levy, Scott Yih and Yashar Mehdad*10:45-12:15 (Regency A & B)
Many NLP tasks require processing long contexts beyond the length limit of pretrained models. In order to scale these models to longer text sequences, many efficient long-range attention variants have been proposed. Despite the abundance of research along this direction, it is still difficult to gauge the relative effectiveness of these models in practical use cases, e.g., if we apply these models following the pretrain-and-finetune paradigm. In this work, we aim to conduct a thorough analysis of these emerging models with large-scale and controlled experiments. For each attention variant, we pretrain large-size models using the same long-doc corpus and then finetune these models for real-world long-context tasks. Our findings reveal pitfalls of an existing widely-used long-range benchmark and show none of the tested efficient attentions can beat a simple local window attention under standard pretraining paradigms. Further analysis on local attention variants suggests that even the commonly used attention-window overlap is not necessary to achieve good downstream results — using disjoint local attentions, we are able to build a simpler and more efficient long-doc QA model that matches the performance of Longformer with half of its pretraining compute.

Reframing Human-AI Collaboration for Generating Free-Text Explanations
Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl and Yejin Choi 10:45-12:15 (Regency A & B)
Large language models are increasingly capable of generating fluent-appearing text with relatively little task-specific supervision. But can these models accurately explain classification decisions? We consider the task of generating free-text explanations using human-written examples in a few-shot manner. We find that (1) authoring higher quality prompts results in higher quality generations; and (2) surprisingly, in a head-to-head comparison, crowdworkers often prefer explanations generated by GPT-3 to crowdsourced explanations in existing datasets. Our human studies also show, however, that while models often produce factual, grammatical, and sufficient explanations, they have room to improve along axes such as providing novel information and supporting the label. We create a pipeline that combines GPT-3 with a supervised filter that incorporates binary acceptability judgments from humans in the loop. Despite the intrinsic subjectivity of acceptability judgments, we demonstrate that acceptability is partially correlated with various fine-grained attributes of explanations. Our approach is able to consistently filter GPT-3-generated explanations deemed acceptable by humans.

Informativeness and Invariance: Two Perspectives on Spurious Correlations in Natural Language
Jacob Eisenstein 10:45-12:15 (Regency A & B)
Spurious correlations are a threat to the trustworthiness of natural language processing systems, motivating research into methods for identifying and eliminating them. However, addressing the problem of spurious correlations requires more clarity on what they are and how they arise in language data. Gardner et al (2021) argue that the compositional nature of language implies that *all* correlations between labels and individual “input features” are spurious. This paper analyzes this proposal in the context of a toy example, demonstrating three distinct conditions that can give rise to feature-label correlations in a simple PCFG. Linking the toy example to a structured causal model shows that (1) feature-label correlations can arise even when the label is invariant to interventions on the feature, and (2) feature-label correlations may be absent even when the label is sensitive to interventions on the feature. Because input features will be individually correlated with labels in all but very rare circumstances, domain knowledge must be applied to identify spurious correlations that pose genuine robustness threats.

On the Diversity and Limits of Human Explanations
Chenhao Tan 10:45-12:15 (Regency A & B)
A growing effort in NLP aims to build datasets of human explanations. However, it remains unclear whether these datasets serve their intended goals. This problem is exacerbated by the fact that the term explanation is overloaded and refers to a broad range of notions with different properties and ramifications. Our goal is to provide an overview of the diversity of explanations, discuss human limitations in providing explanations, and ultimately provide implications for collecting and using human explanations in NLP.

Inspired by prior work in psychology and cognitive sciences, we group existing human explanations in NLP into three categories: proximal mechanism, evidence, and procedure. These three types differ in nature and have implications for the resultant explanations. For instance, procedure is not considered explanation in psychology and connects with a rich body of work on learning from instructions. The diversity of explanations is further evidenced by proxy questions that are needed for annotators to interpret and answer “why is [input] assigned [label]”. Finally, giving explanations may require different, often deeper, understandings than predictions, which casts doubt on whether humans can provide valid explanations in some tasks.

Same Neurons, Different Languages: Probing Morphosyntax in Multilingual Pre-trained Models
Karolina Stanczak, Edoardo Ponti, Lucas Torroba Hennigen, Ryan Cotterell and Isabelle Augenstein 10:45-12:15 (Regency A & B)
The success of multilingual pre-trained models is underpinned by their ability to learn representations shared by multiple languages even in absence of any explicit supervision. However, it remains unclear how these models learn to generalise across languages. In this work, we conjecture that multilingual pre-trained models can derive language-universal abstractions about grammar. In particular, we investigate whether morphosyntactic information is encoded in the same subset of neurons in different languages. We conduct the first large-scale empirical study over 43 languages and 14 morphosyntactic categories with a state-of-the-art neuron-level probe. Our findings show that the cross-lingual

overlap between neurons is significant, but its extent may vary across categories and depends on language proximity and pre-training data size.

[TACL] **Explanation-Based Human Debugging of NLP Models: A Survey**
Piyawat Lertvittayakumjorn and Francesca Toni 10:45-12:15 (Regency A & B)
Debugging a machine learning model is hard since the bug usually involves the training data and the learning process. This becomes even harder for an opaque deep learning model if we have no clue about how the model actually works. In this survey, we review papers that exploit explanations to enable humans to give feedback and debug NLP models. We call this problem explanation-based human debugging (EBHD). In particular, we categorize and discuss existing work along three dimensions of EBHD (the bug context, the workflow, and the experimental setting), compile findings on how EBHD components affect the feedback providers, and highlight open problems that could be future research directions.

Exposing the Limits of Video-Text Models through Contrast Sets
Jae Sung Park, Sheng Shen, Ali Farhadi, Trevor Darrell, Yejin Choi and Anna Rohrbach 10:45-12:15 (Regency A & B)
Recent video-text models can retrieve relevant videos based on text with a high accuracy, but to what extent do they comprehend the semantics of the text? Can they discriminate between similar entities and actions? To answer this, we propose an evaluation framework that probes video-text models with hard negatives. We automatically build contrast sets, where true textual descriptions are manipulated in ways that change their semantics while maintaining plausibility. Specifically, we leverage a pre-trained language model and a set of heuristics to create verb and person entity focused contrast sets. We apply these in the multiple choice video to-text classification setting. We test the robustness of recent methods on the proposed automatic contrast sets, and compare them to additionally collected human-generated counterparts, to assess their effectiveness. We see that model performance suffers across all methods, erasing the gap between recent CLIP-based methods vs. the earlier methods.

FOAM: A Follower-aware Speaker Model For Vision-and-Language Navigation
Zi-Yi Dou and Nanyun Peng 10:45-12:15 (Regency A & B)
The speaker-follower models have proven to be effective in vision-and-language navigation, where a speaker model is used to synthesize new instructions to augment the training data for a follower navigation model. However, in previous work, the speaker model is follower-agnostic and fails to take the state of the follower into consideration. In this paper, we present FOAM, a FOllower-Aware speaker Model that is constantly updated given the follower feedback, so that the generated instructions can be more suitable to the current learning state of the follower. Specifically, we optimize the speaker using a bi-level optimization framework and obtain its training signals by evaluating the follower on labeled data. Experimental results on the Room-to-Room and Room-across-Room datasets demonstrate that our methods can outperform strong baseline models across settings. Analyses also reveal that our generated instructions are of higher quality than the baselines.

MCSE: Multimodal Contrastive Learning of Sentence Embeddings
Miaoran Zhang, Marius Mosbach, David Ifeoluwa Adelani, Michael A. Hedderich and Dietrich Klakow 10:45-12:15 (Regency A & B)
Learning semantically meaningful sentence embeddings is an open problem in natural language processing. In this work, we propose a sentence embedding learning approach that exploits both visual and textual information via a multimodal contrastive objective. Through experiments on a variety of semantic textual similarity tasks, we demonstrate that our approach consistently improves the performance across various datasets and pre-trained encoders. In particular, combining a small amount of multimodal data with a large text-only corpus, we improve the state-of-the-art average Spearman's correlation by 1.7%. By analyzing the properties of the textual embedding space, we show that our model excels in aligning semantically similar sentences, providing an explanation for its improved performance.

Quality-Aware Decoding for Neural Machine Translation
Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. De Souza, Perez Ogayo, Graham Neubig and Andre Martins 10:45-12:15 (Regency A & B)
Despite the progress in machine translation quality estimation and evaluation in the last years, decoding in neural machine translation (NMT) is mostly oblivious to this and centers around finding the most probable translation according to the model (MAP decoding), approximated with beam search. In this paper, we bring together these two lines of research and propose *quality-aware decoding* for NMT, by leveraging recent breakthroughs in reference-free and reference-based MT evaluation through various inference methods like N-best reranking and minimum Bayes risk decoding. We perform an extensive comparison of various possible candidate generation and ranking methods across four datasets and two model classes and find that quality-aware decoding consistently outperforms MAP-based decoding according both to state-of-the-art automatic metrics (COMET and BLEURT) and to human assessments.

Cheat Codes to Quantify Missing Source Information in Neural Machine Translation
Proyag Pal and Kenneth Heafield 10:45-12:15 (Regency A & B)
This paper describes a method to quantify the amount of information $H(t|s)$ added by the target sentence t that is not present in the source s in a neural machine translation system. We do this by providing the model the target sentence in a highly compressed form (a "cheat code"), and exploring the effect of the size of the cheat code. We find that the model is able to capture extra information from just a single float representation of the target and nearly reproduces the target with two 32-bit floats per target token.

Language Model Augmented Monotonic Attention for Simultaneous Translation
Sathish Reddy Indurthi, Mohd Abbas Zaidi, Beomseok Lee, Nikhil Kumar Lakumarapu and Sangha Kim 10:45-12:15 (Regency A & B)
The state-of-the-art adaptive policies for Simultaneous Neural Machine Translation (SNMT) use monotonic attention to perform read/write decisions based on the partial source and target sequences. The lack of sufficient information might cause the monotonic attention to take poor read/write decisions, which in turn negatively affects the performance of the SNMT model. On the other hand, human translators make better read/write decisions since they can anticipate the immediate future words using linguistic information and domain knowledge. In this work, we propose a framework to aid monotonic attention with an external language model to improve its decisions. Experiments on MuST-C English-German and English-French speech-to-text translation tasks show the future information from the language model improves the state-of-the-art monotonic multi-head attention model further.

Building Multilingual Machine Translation Systems That Serve Arbitrary XY Translations
Akiko Eriguchi, Shufang Xie, Tao Qin and Hany Hassan 10:45-12:15 (Regency A & B)
Multilingual Neural Machine Translation (MNMT) enables one system to translate sentences from multiple source languages to multiple target languages, greatly reducing deployment costs compared with conventional bilingual systems. The MNMT training benefit, however, is often limited to many-to-one directions. The model suffers from poor performance in one-to-many and many-to-many with zero-shot setup. To address this issue, this paper discusses how to practically build MNMT systems that serve arbitrary X-Y translation directions while leveraging multilinguality with a two-stage training strategy of pretraining and finetuning. Experimenting with the WMT'21 multilingual translation task, we demonstrate that our systems outperform the conventional baselines of direct bilingual models and pivot translation models for most directions, averagely giving +6.0 and +4.1 BLEU, without the need for architecture change or extra data collection. Moreover,

we also examine our proposed approach in an extremely large-scale data setting to accommodate practical deployment scenarios.

Training Mixed-Domain Translation Models via Federated Learning
Peyman Passban, Tanya Roosta, Rahul Gupta, Ankit Chadha and Clement Chung 10:45-12:15 (Regency A & B)
Training mixed-domain translation models is a complex task that demands tailored architec- tures and costly data preparation techniques. In this work, we leverage federated learning (FL) in order to tackle the problem. Our investiga- tion demonstrates that with slight modifications in the training process, neural machine trans- lation (NMT) engines can be easily adapted when an FL-based aggregation is applied to fuse different domains. Experimental results also show that engines built via FL are able to perform on par with state-of-the-art baselines that rely on centralized training techniques. We evaluate our hypothesis in the presence of five datasets with different sizes, from different domains, to translate from German into English and discuss how FL and NMT can mutually benefit from each other. In addition to provid- ing bench-marking results on the union of FL and NMT, we also propose a novel technique to dynamically control the communication band- width by selecting impactful parameters during FL updates. This is a significant achievement considering the large size of NMT engines that need to be exchanged between FL parties.

Towards Debiasing Translation Artifacts
Koel Dutta Chowdhury, Rricha Jalota, Cristina España-Bonet and Josef Van Genabith 10:45-12:15 (Regency A & B)
Cross-lingual natural language processing relies on translation, either by humans or machines, at different levels, from translating training data to translating test sets. However, compared to original texts in the same language, translations possess distinct qualities referred to as translationese. Previous research has shown that these translation artifacts influence the performance of a variety of cross-lingual tasks. In this work, we propose a novel approach to reducing translationese by extending an established bias-removal technique. We use the Iterative Null-space Projection (INLP) algorithm, and show by measuring classification accuracy before and after debiasing, that translationese is reduced at both sentence and word level. We evaluate the utility of debiasing translationese on a natural language inference (NLI) task, and show that by reducing this bias, NLI accuracy improves. To the best of our knowledge, this is the first study to debias translationese as represented in latent embedding space.

Pretrained Models for Multilingual Federated Learning
Orion Weller, Marc Marone, Vladimir Braverman, Dawn Lawrie and Benjamin Van Durme 10:45-12:15 (Regency A & B)
Since the advent of Federated Learning (FL), research has applied these methods to natural language processing (NLP) tasks. Despite a plethora of papers in FL for NLP, no previous works have studied how multilingual text impacts FL algorithms. Furthermore, multilingual text provides an interesting avenue to examine the impact of non-IID text (e.g. different languages) on FL in naturally occurring data. We explore three multilingual language tasks, language modeling, machine translation, and text classification using differing federated and non-federated learning algorithms. Our results show that using pretrained models reduces the negative effects of FL, helping them to perform near or better than centralized (no privacy) learning, even when using non-IID partitioning.

[CL] **Investigating Language Relationships in Multilingual Sentence Encoders through the Lens of Linguistic Typology**
Rochelle Choenni and Ekaterina Shutova 10:45-12:15 (Regency A & B)
Multilingual sentence encoders have seen much success in cross-lingual model transfer for downstream NLP tasks. The success of this transfer is, however, dependent on the model's ability to encode the patterns of cross-lingual similarity and variation. Yet, we know relatively little about the properties of individual languages or the general patterns of linguistic variation that the models encode. In this article, we investigate these questions by leveraging knowledge from the field of linguistic typology, which studies and documents structural and semantic variation across languages. We propose methods for separating language-specific subspaces within state-of-the-art multilingual sentence encoders (LASER, M-BERT, XLM, and XLM-R) with respect to a range of typological properties pertaining to lexical, morphological, and syntactic structure. Moreover, we investigate how typological information about languages is distributed across all layers of the models. Our results show interesting differences in encoding linguistic variation associated with different pretraining strategies. In addition, we propose a simple method to study how shared typological properties of languages are encoded in two state-of-the-art multilingual models—M-BERT and XLM-R. The results provide insight into their information sharing mechanisms and suggest that these linguistic properties are encoded jointly across typologically similar languages in these models.

DynamicTOC: Persona-based Table of Contents for Consumption of Long Documents
Himanshu Maheshwari, Nethraa Sivakumar, Shelly Jain, Tanvi Karandikar, Vinay Aggarwal, Navita Goyal and Sumit Shekhar 10:45-12:15 (Regency A & B)
Long documents like contracts, financial documents, etc., are often tedious to read through. Linearly consuming (via scrolling or navigation through default table of content) these documents is time-consuming and challenging. These documents are also authored to be consumed by varied entities (referred to as persona in the paper) interested in only certain parts of the document. In this work, we describe DynamicToC, a dynamic table of content-based navigator, to aid in the task of non-linear, persona-based document consumption. DynamicToC highlights sections of interest in the document as per the aspects relevant to different personas. DynamicToC is augmented with short questions to assist the users in understanding underlying content. This uses a novel deep-reinforcement learning technique to generate questions on these persona-clustered paragraphs. Human and automatic evaluations suggest the efficacy of both end-to-end pipeline and different components of DynamicToC.

TWEETSPIN: Fine-grained Propaganda Detection in Social Media Using Multi-View Representations
Prashanth Vijayaraghavan and Soroush Vosoughi 10:45-12:15 (Regency A & B)
Recently, several studies on propaganda detection have involved document and fragment-level analyses of news articles. However, there are significant data and modeling challenges dealing with fine-grained detection of propaganda on social media. In this work, we present TWEETSPIN, a dataset containing tweets that are weakly annotated with different fine-grained propaganda techniques, and propose a neural approach to detect and categorize propaganda tweets across those fine-grained categories. These categories include specific rhetorical and psychological techniques, ranging from leveraging emotions to using logical fallacies. Our model relies on multi-view representations of the input tweet data to (a) extract different aspects of the input text including the context, entities, their relationships, and external knowledge; (b) model their mutual interplay; and (c) effectively speed up the learning process by requiring fewer training examples. Our method allows for representation enrichment leading to better detection and categorization of propaganda on social media. We verify the effectiveness of our proposed method on TWEETSPIN and further probe how the implicit relations between the views impact the performance. Our experiments show that our model is able to outperform several benchmark methods and transfer the knowledge to relatively low-resource news domains.

A Shoulder to Cry on: Towards A Motivational Virtual Assistant for Assuaging Mental Agony
Tulika Saha, Saichethan Miriyala Reddy, Anindya Sundar Das, Sriparna Saha and Pushpak Bhattacharyya 10:45-12:15 (Regency A & B)
Mental Health Disorders continue plaguing humans worldwide. Aggravating this situation is the severe shortage of qualified and competent mental health professionals (MHPs), which underlines the need for developing Virtual Assistants (VAs) that can *assist* MHPs. The data+ML for automation can come from platforms that allow visiting and posting messages in peer-to-peer anonymous manner for sharing their experiences (frequently stigmatized) and seeking support. In this paper, we propose a VA that can act as the first point of contact and comfort for

mental health patients. We curate a dataset, Motivational VA: MotiVAte comprising of 7k dyadic conversations collected from a peer-to-peer support platform. The system employs two mechanisms: (i) Mental Illness Classification: an attention based BERT classifier that outputs the mental disorder category out of the 4 categories, viz., Major Depressive Disorder (MDD), Anxiety, Obsessive Compulsive Disorder (OCD) and Post-traumatic Stress Disorder (PTSD), based on the input ongoing dialog between the support seeker and the VA; and (ii) Mental Illness Conditioned Motivational Dialogue Generation (MI-MDG): a sentiment driven Reinforcement Learning (RL) based motivational response generator. The empirical evaluation demonstrates the system capability by way of outperforming several baselines.

Cross-document Misinformation Detection based on Event Graph Reasoning
Xueqing Wu, Kung-Hsiang Huang, Yi Fung and Heng Ji 10:45-12:15 (Regency A & B)
For emerging events, human readers are often exposed to both real news and fake news. Multiple news articles may contain complementary or contradictory information that readers can leverage to help detect fake news. Inspired by this process, we propose a novel task of cross-document misinformation detection. Given a cluster of topically related news documents, we aim to detect misinformation at both document level and a more fine-grained level, event level. Due to the lack of data, we generate fake news by manipulating real news, and construct 3 new datasets with 422, 276, and 1,413 clusters of topically related documents, respectively. We further propose a graph-based detector that constructs a cross-document knowledge graph using cross-document event coreference resolution and employs a heterogeneous graph neural network to conduct detection at two levels. We then feed the event-level detection results into the document-level detector. Experimental results show that our proposed method significantly outperforms existing methods by up to 7 F1 points on this new task.

A Word is Worth A Thousand Dollars: Adversarial Attack on Tweets Fools Stock Prediction
Yong Xie, Dakuo Wang, Pin-Yu Chen, Jinjun Xiong, Sijia Liu and Oluwasanmi O Koyejo 10:45-12:15 (Regency A & B)
More and more investors and machine learning models rely on social media (e.g., Twitter and Reddit) to gather information and predict movements stock prices. Although text-based models are known to be vulnerable to adversarial attacks, whether stock prediction models have similar vulnerability given necessary constraints is underexplored. In this paper, we experiment with a variety of adversarial attack configurations to fool three stock prediction victim models. We address the task of adversarial generation by solving combinatorial optimization problems with semantics and budget constraints. Our results show that the proposed attack method can achieve consistent success rates and cause significant monetary loss in trading simulation by simply concatenating a perturbed but semantically similar tweet.

Session 2 - 14:30-16:00

Interpretability

14:30-16:00 (Columbia A)

ElitePLM: An Empirical Study on General Language Ability Evaluation of Pretrained Language Models
Junyi Li, Tianyi Tang, Zheng Gong, Lixin Yang, Zhuohao Yu, Zhipeng Chen, Jingyuan Wang, Xin Zhao and Ji-Rong Wen 14:30-14:45 (Columbia A)
Nowadays, pretrained language models (PLMs) have dominated the majority of NLP tasks. While, little research has been conducted on systematically evaluating the language abilities of PLMs. In this paper, we present a large-scale empirical study on general language ability evaluation of PLMs (ElitePLM). In our study, we design four evaluation dimensions, i.e. memory, comprehension, reasoning, and composition, to measure ten widely-used PLMs within five categories. Our empirical results demonstrate that: (1) PLMs with varying training objectives and strategies are good at different ability tests; (2) fine-tuning PLMs in downstream tasks is usually sensitive to the data size and distribution; (3) PLMs have excellent transferability between similar tasks. Moreover, the prediction results of PLMs in our experiments are released as an open resource for more deep and detailed analysis on the language abilities of PLMs. This paper can guide the future work to select, apply, and design PLMs for specific tasks. We have made all the details of experiments publicly available at `https://github.com/RUCAIBox/ElitePLM`.

When Does Syntax Mediate Neural Language Model Performance? Evidence from Dropout Probes
Mycal Tucker, Tiwalayo Eisape, Peng Qian, Roger P. Levy and Julie Shah 14:45-15:00 (Columbia A)
Recent causal probing literature reveals when language models and syntactic probes use similar representations. Such techniques may yield “false negative” causality results: models may use representations of syntax, but probes may have learned to use redundant encodings of the same syntactic information. We demonstrate that models do encode syntactic information redundantly and introduce a new probe design that guides probes to consider all syntactic information present in embeddings. Using these probes, we find evidence for the use of syntax in models where prior methods did not, allowing us to boost model performance by injecting syntactic information into representations.

ExSum: From Local Explanations to Model Understanding
Yilun Zhou, Marco Tulio Ribeiro and Julie Shah 15:00-15:15 (Columbia A)
Interpretability methods are developed to understand the working mechanisms of black-box models, which is crucial to their responsible deployment. Fulfilling this goal requires both that the explanations generated by these methods are correct and that people can easily and reliably understand them. While the former has been addressed in prior work, the latter is often overlooked, resulting in informal model understanding derived from a handful of local explanations. In this paper, we introduce explanation summary (ExSum), a mathematical framework for quantifying model understanding, and propose metrics for its quality assessment. On two domains, ExSum highlights various limitations in the current practice, helps develop accurate model understanding, and reveals easily overlooked properties of the model. We also connect understandability to other properties of explanations such as human alignment, robustness, and counterfactual similarity and plausibility.

Is "My Favorite New Movie" My Favorite Movie? Probing the Understanding of Recursive Noun Phrases
Qing Lyu, Zheng Hua, Daoxin Li, Li Zhang, Marianna Apidianaki and Chris Callison-Burch 15:15-15:30 (Columbia A)
Recursive noun phrases (NPs) have interesting semantic properties. For example, "my favorite new movie" is not necessarily my favorite movie, whereas "my new favorite movie" is. This is common sense to humans, yet it is unknown whether language models have such knowledge. We introduce the Recursive Noun Phrase Challenge (RNPC), a dataset of three textual inference tasks involving textual entailment and event plausibility comparison, precisely targeting the understanding of recursive NPs. When evaluated on RNPC, state-of-the-art Transformer models only perform around chance. Still, we show that such knowledge is learnable with appropriate data. We further probe the models for relevant linguistic features that can be learned from our tasks, including modifier semantic category and modifier scope. Finally, models trained on RNPC achieve strong zero-shot performance on an extrinsic Harm Detection evaluation task, showing the usefulness of the understanding of recursive NPs in downstream applications.

Lifelong Pretraining: Continually Adapting Language Models to Emerging Corpora
Xisen Jin, Dejiao Zhang, Henghui Zhu, Wei Xiao, Shang-Wen Li, Xiaokai Wei, Andrew Arnold and Xiang Ren 15:30-15:45 (Columbia A)
Pretrained language models (PTLMs) are typically learned over a large, static corpus and further fine-tuned for various downstream tasks. However, when deployed in the real world, a PTLM-based model must deal with data distributions that deviates from what the PTLM was initially trained on. In this paper, we study a lifelong language model pretraining challenge where a PTLM is continually updated so as to adapt to emerging data. Over a domain-incremental research paper stream and a chronologically-ordered tweet stream, we incrementally pretrain a PTLM with different continual learning algorithms, and keep track of the downstream task performance (after fine-tuning). We evaluate PTLM's ability to adapt to new corpora while retaining learned knowledge in earlier corpora. Our experiments show distillation-based approaches to be most effective in retaining downstream performance in earlier domains. The algorithms also improve knowledge transfer, allowing models to achieve better downstream performance over latest data, and improve temporal generalization when distribution gaps exist between training and evaluation because of time. We believe our problem formulation, methods, and analysis will inspire future studies towards continual pretraining of language models.

Even the Simplest Baseline Needs Careful Re-investigation: A Case Study on XML-CNN
Si-An Chen, Jie-jyun Liu, Tsung-Han Yang, Hsuan-Tien Lin and Chih-Jen Lin 15:45-16:00 (Columbia A)
The power and the potential of deep learning models attract many researchers to design advanced and sophisticated architectures. Nevertheless, the progress is sometimes unreal due to various possible reasons. In this work, through an astonishing example we argue that more efforts should be paid to ensure the progress in developing a new deep learning method. For a highly influential multi-label text classification method XML-CNN, we show that the superior performance claimed in the original paper was mainly due to some unbelievable coincidences. We re-examine XML-CNN and make a re-implementation which reveals some contradictory findings to the claims in the original paper. Our study suggests suitable baselines for multi-label text classification tasks and confirms that the progress on a new architecture cannot be confidently justified without a cautious investigation.

Semantics

14:30-16:00 (Columbia C)

Testing the Ability of Language Models to Interpret Figurative Language
Emmy Liu, Chenxuan Cui, Kenneth Zheng and Graham Neubig 14:30-14:45 (Columbia C)
Figurative and metaphorical language are commonplace in discourse, and figurative expressions play an important role in communication and cognition. However, figurative language has been a relatively under-studied area in NLP, and it remains an open question to what extent modern language models can interpret nonliteral phrases. To address this question, we introduce Fig-QA, a Winograd-style nonliteral language understanding task consisting of correctly interpreting paired figurative phrases with divergent meanings. We evaluate the performance of several state-of-the-art language models on this task, and find that although language models achieve performance significantly over chance, they still fall short of human performance, particularly in zero- or few-shot settings. This suggests that further work is needed to improve the nonliteral reasoning capabilities of language models.

Compositional Task-Oriented Parsing as Abstractive Question Answering
Wenting Zhao, Konstantine Arkoudas, Weiqi Sun and Claire Cardie 14:45-15:00 (Columbia C)
Task-oriented parsing (TOP) aims to convert natural language into machine-readable representations of specific tasks, such as setting an alarm. A popular approach to TOP is to apply seq2seq models to generate linearized parse trees. A more recent line of work argues that pretrained seq2seq2 models are better at generating outputs that are themselves natural language, so they replace linearized parse trees with canonical natural-language paraphrases that can then be easily translated into parse trees, resulting in so-called naturalized parsers. In this work we continue to explore naturalized semantic parsing by presenting a general reduction of TOP to abstractive question answering that overcomes some limitations of canonical paraphrasing. Experimental results show that our QA-based technique outperforms state-of-the-art methods in full-data settings while achieving dramatic improvements in few-shot settings.

Linguistic Frameworks Go Toe-to-Toe at Neuro-Symbolic Language Modeling
Jakob Prange, Nathan Schneider and Lingpeng Kong 15:00-15:15 (Columbia C)
We examine the extent to which, in principle, different syntactic and semantic graph representations can complement and improve neural language modeling. Specifically, by conditioning on a subgraph encapsulating the locally relevant sentence history, can a model make better next-word predictions than a pretrained sequential language model alone? With an ensemble setup consisting of GPT-2 and ground-truth graphs from one of 7 different formalisms, we find that the graph information indeed improves perplexity and other metrics. Moreover, this architecture provides a new way to compare different frameworks of linguistic representation. In our oracle graph setup, training and evaluating on English WSJ, semantic constituency structures prove most useful to language modeling performance—outpacing syntactic constituency structures as well as syntactic and semantic dependency structures.

Improving Compositional Generalization with Latent Structure and Data Augmentation
Linlu Qiu, Peter Shaw, Panupong Pasupat, Pawel Krzysztof Nowak, Tal Linzen, Fei Sha and Kristina Toutanova 15:15-15:30 (Columbia C)
Generic unstructured neural networks have been shown to struggle on out-of-distribution compositional generalization. Compositional data augmentation via example recombination has transferred some prior knowledge about compositionality to such black-box neural models for several semantic parsing tasks, but this often required task-specific engineering or provided limited gains. We present a more powerful data recombination method using a model called Compositional Structure Learner (CSL). CSL is a generative model with a quasi-synchronous context-free grammar backbone, which we induce from the training data. We sample recombined examples from CSL and add them to the fine-tuning data of a pre-trained sequence-to-sequence model (T5). This procedure effectively transfers most of CSL's compositional bias to T5 for diagnostic tasks, and results in a model even stronger than a T5-CSL ensemble on two real world compositional generalization tasks. This results in new state-of-the-art performance for these challenging semantic parsing tasks requiring generalization to both natural language variation and novel compositions of elements.

DiffCSE: Difference-based Contrastive Learning for Sentence Embeddings
Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljacic, Shang-Wen Li, Scott Yih, Yoon Kim and James R. Glass 15:30-15:45 (Columbia C)
We propose DiffCSE, an unsupervised contrastive learning framework for learning sentence embeddings. DiffCSE learns sentence embeddings that are sensitive to the difference between the original sentence and an edited sentence, where the edited sentence is obtained by stochastically masking out the original sentence and then sampling from a masked language model. We show that DiffSCE is an instance of equivariant contrastive learning, which generalizes contrastive learning and learns representations that are insensitive to certain types of

augmentations and sensitive to other "harmful" types of augmentations. Our experiments show that DiffCSE achieves state-of-the-art results among unsupervised sentence representation learning methods, outperforming unsupervised SimCSE by 2.3 absolute points on semantic textual similarity tasks.

Bilingual Tabular Inference: A Case Study on Indic Languages
Chaitanya Agarwal, Vivek Gupta, Anoop Kunchukuttan and Manish Shrivastava 15:45-16:00 (Columbia C)
Existing research on Tabular Natural Language Inference (TNLI) exclusively examines the task in a monolingual setting where the tabular premise and hypothesis are in the same language. However, due to the uneven distribution of text resources on the web across languages, it is common to have the tabular premise in a high resource language and the hypothesis in a low resource language. As a result, we present the challenging task of bilingual Tabular Natural Language Inference (bTNLI), in which the tabular premise and a hypothesis over it are in two separate languages. We construct EI-InfoTabS: an English-Indic bTNLI dataset by translating the textual hypotheses of the English TNLI dataset InfoTabS into eleven major Indian languages. We thoroughly investigate how pre-trained multilingual models learn and perform on EI-InfoTabS. Our study shows that the performance on bTNLI can be close to its monolingual counterpart, with translate-train, translate-test and unified-train being strongly competitive baselines.

Language Resources & Evaluation 1

14:30-16:00 (Columbia D)

Bidimensional Leaderboards: Generate and Evaluate Language Hand in Hand
Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Lavinia Dunagan, Jacob Daniel Morrison, Alexander Fabbri, Yejin Choi and Noah Smith
14:30-14:45 (Columbia D)
Natural language processing researchers have identified limitations of evaluation methodology for generation tasks, with new questions raised about the validity of automatic metrics and of crowdworker judgments. Meanwhile, efforts to improve generation models tend to depend on simple n-gram overlap metrics (e.g., BLEU, ROUGE). We argue that new advances on models and metrics should each more directly benefit and inform the other. We therefore propose a generalization of leaderboards, bidimensional leaderboards (Billboards), that simultaneously tracks progress in language generation models and metrics for their evaluation. Unlike conventional unidimensional leaderboards that sort submitted systems by predetermined metrics, a Billboard accepts both generators and evaluation metrics as competing entries. A Billboard automatically creates an ensemble metric that selects and linearly combines a few metrics based on a global analysis across generators. Further, metrics are ranked based on their correlation with human judgments. We release four Billboards for machine translation, summarization, and image captioning. We demonstrate that a linear ensemble of a few diverse metrics sometimes substantially outperforms existing metrics in isolation. Our mixed-effects model analysis shows that most automatic metrics, especially the reference-based ones, overrate machine over human generation, demonstrating the importance of updating metrics as generation models become stronger (and perhaps more similar to humans) in the future.

CORWA: A Citation-Oriented Related Work Annotation Dataset
Xiangci Li, Biswadip Mandal and Jessica Ouyang 14:45-15:00 (Columbia D)
Academic research is an exploratory activity to discover new solutions to problems. By this nature, academic research works perform literature reviews to distinguish their novelties from prior work. In natural language processing, this literature review is usually conducted under the “Related Work” section. The task of related work generation aims to automatically generate the related work section given the rest of the research paper and a list of papers to cite. Prior work on this task has focused on the sentence as the basic unit of generation, neglecting the fact that related work sections consist of variable length text fragments derived from different information sources. As a first step toward a linguistically-motivated related work generation framework, we present a Citation Oriented Related Work Annotation (CORWA) dataset that labels different types of citation text fragments from different information sources. We train a strong baseline model that automatically tags the CORWA labels on massive unlabeled related work section texts. We further suggest a novel framework for human-in-the-loop, iterative, abstractive related work generation.

Shedding New Light on the Language of the Dark Web
Youngjin Jin, Eugene Jang, Yongjae Lee, Seungwon Shin and Jin-Woo Chung 15:00-15:15 (Columbia D)
The hidden nature and the limited accessibility of the Dark Web, combined with the lack of public datasets in this domain, make it difficult to study its inherent characteristics such as linguistic properties. Previous works on text classification of Dark Web domain have suggested that the use of deep neural models may be ineffective, potentially due to the linguistic differences between the Dark and Surface Webs. However, not much work has been done to uncover the linguistic characteristics of the Dark Web. This paper introduces CoDA, a publicly available Dark Web dataset consisting of 10000 web documents tailored towards text-based Dark Web analysis. By leveraging CoDA, we conduct a thorough linguistic analysis of the Dark Web and examine the textual differences between the Dark Web and the Surface Web. We also assess the performance of various methods of Dark Web page classification. Finally, we compare CoDA with an existing public Dark Web dataset and evaluate their suitability for various use cases.

A Few Thousand Translations Go a Long Way! Leveraging Pre-trained Models for African News Translation
David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Chinenye Emezue, Colin Leong, Michael Beukman, Shamsuddeen Hassan Muhammad, Guyo Dub Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ayoade Ajibade, Tunde Oluwaseyi Ajayi, Yvonne Wambui Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Koffi Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula and Sam Manthalu 15:15-15:30 (Columbia D)
Recent advances in the pre-training for language models leverage large-scale datasets to create multilingual models. However, low-resource languages are mostly left out in these datasets. This is primarily because many widely spoken languages that are not well represented on the web and therefore excluded from the large-scale crawls for datasets. Furthermore, downstream users of these models are restricted to the selection of languages originally chosen for pre-training. This work investigates how to optimally leverage existing pre-trained models to create low-resource translation systems for 16 African languages. We focus on two questions: 1) How can pre-trained models be used for languages not included in the initial pretraining? and 2) How can the resulting translation models effectively transfer to new domains? To answer these questions, we create a novel African news corpus covering 16 languages, of which eight languages are not part of any existing evaluation dataset. We demonstrate that the most effective strategy for transferring both additional languages and additional domains is to leverage small quantities of high-quality translation data to fine-tune large pre-trained models.

Does Summary Evaluation Survive Translation to Other Languages?
Spencer Braun, Oleg Vasilyev, Neslihan Iskender and John Bohannon 15:30-15:45 (Columbia D)
The creation of a quality summarization dataset is an expensive, time-consuming effort, requiring the production and evaluation of summaries by both trained humans and machines. The returns to such an effort would increase significantly if the dataset could be used in additional languages without repeating human annotations. To investigate how much we can trust machine translation of summarization datasets, we translate the English SummEval dataset to seven languages and compare performances across automatic evaluation measures. We explore equivalence testing as the appropriate statistical paradigm for evaluating correlations between human and automated scoring of summaries. We also consider the effect of translation on the relative performance between measures. We find some potential for dataset reuse in languages similar to the source and along particular dimensions of summary quality. Our code and data can be found at https://github.com/PrimerAI/primer-research/.

DISAPERE: A Dataset for Discourse Structure in Peer Review Discussions
Neha Nayak Kennard, Tim O'Gorman, Rajarshi Das, Akshay Sharma, Chhandak Bagchi, Matthew Clinton, Pranay Kumar Yelugam, Hamed Zamani and Andrew McCallum 15:45-16:00 (Columbia D)
At the foundation of scientific evaluation is the labor-intensive process of peer review. This critical task requires participants to consume vast amounts of highly technical text. Prior work has annotated different aspects of review argumentation, but discourse relations between reviews and rebuttals have yet to be examined. We present DISAPERE, a labeled dataset of 20k sentences contained in 506 review-rebuttal pairs in English, annotated by experts. DISAPERE synthesizes label sets from prior work and extends them to include fine-grained annotation of the rebuttal sentences, characterizing their context in the review and the authors' stance towards review arguments. Further, we annotate *every* review and rebuttal sentence. We show that discourse cues from rebuttals can shed light on the quality and interpretation of reviews. Further, an understanding of the argumentative strategies employed by the reviewers and authors provides useful signal for area chairs and other decision makers.

Machine Translation 1

14:30-16:00 (Elwha A)

Original or Translated? A Causal Analysis of the Impact of Translationese on Machine Translation Performance
Jingwei Ni, Zhijing Jin, Markus Freitag, Mrinmaya Sachan and Bernhard Schölkopf 14:30-14:45 (Elwha A)
Human-translated text displays distinct features from naturally written text in the same language. This phenomena, known as translationese, has been argued to confound the machine translation (MT) evaluation. Yet, we find that existing work on translationese neglects some important factors and the conclusions are mostly correlational but not causal. In this work, we collect CausalMT, a dataset where the MT training data are also labeled with the human translation directions. We inspect two critical factors, the train-test direction match (whether the human translation directions in the training and test sets are aligned), and data-model direction match (whether the model learns in the same direction as the human translation direction in the dataset). We show that these two factors have a large causal effect on the MT performance, in addition to the test-model direction mismatch highlighted by existing work on the impact of translationese. In light of our findings, we provide a set of suggestions for MT training and evaluation. Our code and data are at https://github.com/EdisonNi-hku/CausalMT

On Systematic Style Differences between Unsupervised and Supervised MT and an Application for High-Resource Machine Translation
Kelly Marchisio, Markus Freitag and David Grangier 14:45-15:00 (Elwha A)
Modern unsupervised machine translation (MT) systems reach reasonable translation quality under clean and controlled data conditions. As the performance gap between supervised and unsupervised MT narrows, it is interesting to ask whether the different training methods result in systematically different output beyond what is visible via quality metrics like adequacy or BLEU. We compare translations from supervised and unsupervised MT systems of similar quality, finding that unsupervised output is more fluent and more structurally different in comparison to human translation than is supervised MT. We then demonstrate a way to combine the benefits of both methods into a single system which results in improved adequacy and fluency as rated by human evaluators. Our results open the door to interesting discussions about how supervised and unsupervised MT might be different yet mutually-beneficial.

The Devil is in the Details: On the Pitfalls of Vocabulary Selection in Neural Machine Translation
Tobias Domhan, Eva Hasler, Ke Tran, Sony Trenous, Bill Byrne and Felix Hieber 15:00-15:15 (Elwha A)
Vocabulary selection, or lexical shortlisting, is a well-known technique to improve latency of Neural Machine Translation models by constraining the set of allowed output words during inference. The chosen set is typically determined by separately trained alignment model parameters, independent of the source-sentence context at inference time. While vocabulary selection appears competitive with respect to automatic quality metrics in prior work, we show that it can fail to select the right set of output words, particularly for semantically non-compositional linguistic phenomena such as idiomatic expressions, leading to reduced translation quality as perceived by humans. Trading off latency for quality by increasing the size of the allowed set is often not an option in real-world scenarios. We propose a model of vocabulary selection, integrated into the neural translation model, that predicts the set of allowed output words from contextualized encoder representations. This restores translation quality of an unconstrained system, as measured by human evaluations on WMT newstest2020 and idiomatic expressions, at an inference latency competitive with alignment-based selection using aggressive thresholds, thereby removing the dependency on separately trained alignment models.

BlonDe: An Automatic Evaluation Metric for Document-level Machine Translation
Yuchen Eleanor Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan and Ming Zhou 15:15-15:30 (Elwha A)
Standard automatic metrics, e.g. BLEU, are not reliable for document-level MT evaluation. They can neither distinguish document-level improvements in translation quality from sentence-level ones, nor identify the discourse phenomena that cause context-agnostic translations. This paper introduces a novel automatic metric BlonDe to widen the scope of automatic MT evaluation from sentence to document level. BlonDe takes discourse coherence into consideration by categorizing discourse-related spans and calculating the similarity-based F1 measure of categorized spans. We conduct extensive comparisons on a newly constructed dataset BWB. The experimental results show that BlonDe possesses better selectivity and interpretability at the document-level, and is more sensitive to document-level nuances. In a large-scale human study, BlonDe also achieves significantly higher Pearson's r correlation with human judgments compared to previous metrics.

Non-Autoregressive Machine Translation: It's Not as Fast as it Seems
Jindřich Helcl, Barry Haddow and Alexandra Birch 15:30-15:45 (Elwha A)
Efficient machine translation models are commercially important as they can increase inference speeds, and reduce costs and carbon emis-

sions. Recently, there has been much interest in non-autoregressive (NAR) models, which promise faster translation. In parallel to the research on NAR models, there have been successful attempts to create optimized autoregressive models as part of the WMT shared task on efficient translation. In this paper, we point out flaws in the evaluation methodology present in the literature on NAR models and we provide a fair comparison between a state-of-the-art NAR model and the autoregressive submissions to the shared task. We make the case for consistent evaluation of NAR models, and also for the importance of comparing NAR models with other widely used methods for improving efficiency. We run experiments with a connectionist-temporal-classification-based (CTC) NAR model implemented in C++ and compare it with AR models using wall clock times. Our results show that, although NAR models are faster on GPUs, with small batch sizes, they are almost always slower under more realistic usage conditions. We call for more realistic and extensive evaluation of NAR models in future work.

[TACL] **High Quality Rather than High Model Probability: Minimum Bayes Risk Decoding with Neural Metrics**
Markus Freitag, David Grangier, Qijun Tan and Bowen Liang 15:45-16:00 (Elwha A)
This work applies Minimum Bayes Risk (MBR) decoding on unbiased samples to optimize diverse automated metrics of translation quality. Automatic metrics in machine translation have made tremendous progress recently. In particular, neural metrics, fine-tuned on human ratings (e.g. bleurt, or comet) are outperforming surface metrics in terms of correlations to human judgements. Our experiments show that the combination of a neural translation model with a neural reference-based metric, bleurt, results in significant improvement in automatic and human evaluations. This improvement is obtained with translations different from classical beam-search output: these translations have much lower likelihood and are less favored by surface metrics like bleu.

NLP Applications 1

14:30-16:00 (Elwha B)

ScAN: Suicide Attempt and Ideation Events Dataset
Bhanu Pratap Singh Rawat, Samuel Kovaly, Hong Yu and Wilfred Pigeon 14:30-14:45 (Elwha B)
Suicide is an important public health concern and one of the leading causes of death worldwide. Suicidal behaviors, including suicide attempts (SA) and suicide ideations (SI), are leading risk factors for death by suicide. Information related to patients' previous and current SA and SI are frequently documented in the electronic health record (EHR) notes. Accurate detection of such documentation may help improve surveillance and predictions of patients' suicidal behaviors and alert medical professionals for suicide prevention efforts. In this study, we first built Suicide Attempt and Ideation Events (ScAN) dataset, a subset of the publicly available MIMIC III dataset spanning over 12k+ EHR notes with 19k+ annotated SA and SI events information. The annotations also contain attributes such as method of suicide attempt. We also provide a strong baseline model ScANER (Suicide Attempt and Ideation Events Retriever), a multi-task RoBERTa-based model with a retrieval module to extract all the relevant suicidal behavioral evidences from EHR notes of an hospital-stay and, and a prediction module to identify the type of suicidal behavior (SA and SI) concluded during the patient's stay at the hospital. ScANER achieved a macro-weighted F1-score of 0.83 for identifying suicidal behavioral evidences and a macro F1-score of 0.78 and 0.60 for classification of SA and SI for the patient's hospital-stay, respectively. ScAN and ScANER are publicly available.

DUCK: Rumour Detection on Social Media by Modelling User and Comment Propagation Networks
Lin Tian, Xiuzhen Zhang and Jey Han Lau 14:45-15:00 (Elwha B)
Social media rumours, a form of misinformation, can mislead the public and cause significant economic and social disruption. Motivated by the observation that the user network — which captures *who* engage with a story — and the comment network — which captures *how* they react to it — provide complementary signals for rumour detection, in this paper, we propose DUCK (rumour *d*etection with *u*ser and *c*omment networ*k*s) for rumour detection on social media. We study how to leverage transformers and graph attention networks to jointly model the contents and structure of social media conversations, as well as the network of users who engaged in these conversations. Over four widely used benchmark rumour datasets in English and Chinese, we show that DUCK produces superior performance for detecting rumours, creating a new state-of-the-art. Source code for DUCK is available at: https://github.com/l tian678/DUCK-code.

Early Rumor Detection Using Neural Hawkes Process with a New Benchmark Dataset
Fengzhu Zeng and Wei Gao 15:00-15:15 (Elwha B)
Little attention has been paid on EArly Rumor Detection (EARD), and EARD performance was evaluated inappropriately on a few datasets where the actual early-stage information is largely missing. To reverse such situation, we construct BEARD, a new Benchmark dataset for EARD, based on claims from fact-checking websites by trying to gather as many early relevant posts as possible. We also propose HEARD, a novel model based on neural Hawkes process for EARD, which can guide a generic rumor detection model to make timely, accurate and stable predictions. Experiments show that HEARD achieves effective EARD performance on two commonly used general rumor detection datasets and our BEARD dataset.

Frustratingly Easy System Combination for Grammatical Error Correction
Muhammad Reza Qorib, Seung-Hoon Na and Hwee Tou Ng 15:15-15:30 (Elwha B)
In this paper, we formulate system combination for grammatical error correction (GEC) as a simple machine learning task: binary classification. We demonstrate that with the right problem formulation, a simple logistic regression algorithm can be highly effective for combining GEC models. Our method successfully increases the F0.5 score from the highest base GEC system by 4.2 points on the CoNLL-2014 test set and 7.2 points on the BEA-2019 test set. Furthermore, our method outperforms the state of the art by 4.0 points on the BEA-2019 test set, 1.2 points on the CoNLL-2014 test set with original annotation, and 3.4 points on the CoNLL-2014 test set with alternative annotation. We also show that our system combination generates better corrections with higher F0.5 scores than the conventional ensemble.

On the Use of Bert for Automated Essay Scoring: Joint Learning of Multi-Scale Essay Representation
Yongjie Wang, Chuang Wang, Ruobing Li and Hui Lin 15:30-15:45 (Elwha B)
In recent years, pre-trained models have become dominant in most natural language processing (NLP) tasks. However, in the area of Automated Essay Scoring (AES), pre-trained models such as BERT have not been properly used to outperform other deep learning models such as LSTM. In this paper, we introduce a novel multi-scale essay representation for BERT that can be jointly learned. We also employ multiple losses and transfer learning from out-of-domain essays to further improve the performance. Experiment results show that our approach derives much benefit from joint learning of multi-scale essay representation and obtains almost the state-of-the-art result among all deep learning models in the ASAP task. Our multi-scale essay representation also generalizes well to CommonLit Readability Prize data set, which suggests that the novel text representation proposed in this paper may be a new and effective choice for long-text tasks.

KCD: Knowledge Walks and Textual Cues Enhanced Political Perspective Detection in News Media
Wenqian Zhang, Shangbin Feng, Zilong Chen, Zhenyu Lei, Jundong Li and Minnan Luo 15:45-16:00 (Elwha B)

Political perspective detection has become an increasingly important task that can help combat echo chambers and political polarization. Previous approaches generally focus on leveraging textual content to identify stances, while they fail to reason with background knowledge or leverage the rich semantic and syntactic textual labels in news articles. In light of these limitations, we propose KCD, a political perspective detection approach to enable multi-hop knowledge reasoning and incorporate textual cues as paragraph-level labels. Specifically, we firstly generate random walks on external knowledge graphs and infuse them with news text representations. We then construct a heterogeneous information network to jointly model news content as well as semantic, syntactic and entity cues in news articles. Finally, we adopt relational graph neural networks for graph-level representation learning and conduct political perspective detection. Extensive experiments demonstrate that our approach outperforms state-of-the-art methods on two benchmark datasets. We further examine the effect of knowledge walks and textual cues and how they contribute to our approach's data efficiency.

In-person Poster Session 2

14:30-16:00 (Regency A & B)

Towards a Progression-Aware Autonomous Dialogue Agent
Abraham Sanders, Tomek Strzalkowski, Mei Si, Albert Chang, Deepanshu Dey, Jonas Braasch and Dakuo Wang 14:30-16:00 (Regency A & B)
Recent advances in large-scale language modeling and generation have enabled the creation of dialogue agents that exhibit human-like responses in a wide range of conversational scenarios spanning a diverse set of tasks, from general chit-chat to focused goal-oriented discourse. While these agents excel at generating high-quality responses that are relevant to prior context, they suffer from a lack of awareness of the overall direction in which the conversation is headed, and the likelihood of task success inherent therein. Thus, we propose a framework in which dialogue agents can evaluate the progression of a conversation toward or away from desired outcomes, and use this signal to inform planning for subsequent responses. Our framework is composed of three key elements: (1) the notion of a "global" dialogue state (GDS) space, (2) a task-specific progression function (PF) computed in terms of a conversation's trajectory through this space, and (3) a planning mechanism based on dialogue rollouts by which an agent may use progression signals to select its next response.

Building a Role Specified Open-Domain Dialogue System Leveraging Large-Scale Language Models
Sanghwan Bae, Donghyun Kwak, Sungdong Kim, Donghoon Ham, Soyoung Kang, Sang-Woo Lee and Woomyoung Park 14:30-16:00 (Regency A & B)
Recent open-domain dialogue models have brought numerous breakthroughs. However, building a chat system is not scalable since it often requires a considerable volume of human-human dialogue data, especially when enforcing features such as persona, style, or safety. In this work, we study the challenge of imposing roles on open-domain dialogue systems, with the goal of making the systems maintain consistent roles while conversing naturally with humans. To accomplish this, the system must satisfy a role specification that includes certain conditions on the stated features as well as a system policy on whether or not certain types of utterances are allowed. For this, we propose an efficient data collection framework leveraging in-context few-shot learning of large-scale language models for building role-satisfying dialogue dataset from scratch. We then compare various architectures for open-domain dialogue systems in terms of meeting role specifications while maintaining conversational abilities. Automatic and human evaluations show that our models return few out-of-bounds utterances, keeping competitive performance on general metrics. We release a Korean dialogue dataset we built for further research.

Emp-RFT: Empathetic Response Generation via Recognizing Feature Transitions between Utterances
Wongyu Kim, Youbin Ahn, Donghyun Kim and Kyong-Ho Lee 14:30-16:00 (Regency A & B)
Each utterance in multi-turn empathetic dialogues has features such as emotion, keywords, and utterance-level meaning. Feature transitions between utterances occur naturally. However, existing approaches fail to perceive the transitions because they extract features for the context at the coarse-grained level. To solve the above issue, we propose a novel approach of recognizing feature transitions between utterances, which helps understand the dialogue flow and better grasp the features of utterance that needs attention. Also, we introduce a response generation strategy to help focus on emotion and keywords related to appropriate features when generating responses. Experimental results show that our approach outperforms baselines and especially, achieves significant improvements on multi-turn dialogues.

Database Search Results Disambiguation for Task-Oriented Dialog Systems
Kun Qian, Satwik Kottur, Ahmad Beirami, Shahin Shayandeh, Paul A. Crook, Alborz Geramifard, Zhou Yu and Chinnadhurai Sankar 14:30-16:00 (Regency A & B)
As task-oriented dialog systems are becoming increasingly popular in our lives, more realistic tasks have been proposed and explored. However, new practical challenges arise. For instance, current dialog systems cannot effectively handle multiple search results when querying a database, due to the lack of such scenarios in existing public datasets. In this paper, we propose Database Search Result (DSR) Disambiguation, a novel task that focuses on disambiguating database search results, which enhances user experience by allowing them to choose from multiple options instead of just one. To study this task, we augment the popular task-oriented dialog datasets (MultiWOZ and SGD) with turns that resolve ambiguities by (a) synthetically generating turns through a pre-defined grammar, and (b) collecting human paraphrases for a subset. We find that training on our augmented dialog data improves the model's ability to deal with ambiguous scenarios, without sacrificing performance on unmodified turns. Furthermore, pre-fine tuning and multi-task learning help our model to improve performance on DSR-disambiguation even in the absence of in-domain data, suggesting that it can be learned as a universal dialog skill. Our data and code will be made publicly available.

Learning Dialogue Representations from Consecutive Utterances
Zhihan Zhou, Dejiao Zhang, Wei Xiao, Nicholas Dingwall, Xiaofei Ma, Andrew Arnold and Bing Xiang 14:30-16:00 (Regency A & B)
Learning high-quality dialogue representations is essential for solving a variety of dialogue-oriented tasks, especially considering that dialogue systems often suffer from data scarcity. In this paper, we introduce Dialogue Sentence Embedding (DSE), a self-supervised contrastive learning method that learns effective dialogue representations suitable for a wide range of dialogue tasks. DSE learns from dialogues by taking consecutive utterances of the same dialogue as positive pairs for contrastive learning. Despite its simplicity, DSE achieves significantly better representation capability than other dialogue representation and universal sentence representation models. We evaluate DSE on five downstream dialogue tasks that examine dialogue representation at different semantic granularities. Experiments in few-shot and zero-shot settings show that DSE outperforms baselines by a large margin, for example, it achieves 13% average performance improvement over the strongest unsupervised baseline in 1-shot intent classification on 6 datasets. We also provide analyses on the benefits and limitations of our model.

Knowledge-Grounded Dialogue Generation with a Unified Knowledge Representation
Yu Li, Baolin Peng, Yelong Shen, Yi Mao, Lars Liden, Zhou Yu and Jianfeng Gao 14:30-16:00 (Regency A & B)

Knowledge-grounded dialogue systems are challenging to build due to the lack of training data and heterogeneous knowledge sources. Existing systems perform poorly on unseen topics due to limited topics covered in the training data. In addition, it is challenging to generalize to the domains that require different types of knowledge sources. To address the above challenges, we present PLUG, a language model that homogenizes different knowledge sources to a unified knowledge representation for knowledge-grounded dialogue generation tasks. We first retrieve relevant information from heterogeneous knowledge sources (e.g., wiki, dictionary, or knowledge graph); Then the retrieved knowledge is transformed into text and concatenated with dialogue history to feed into the language model for generating responses. PLUG is pre-trained on a large-scale knowledge-grounded dialogue corpus. The empirical evaluation on two benchmarks shows that PLUG generalizes well across different knowledge-grounded dialogue tasks. It achieves comparable performance with state-of-the-art methods in the fully-supervised setting and significantly outperforms other approaches in zero-shot and few-shot settings.

On the Origin of Hallucinations in Conversational Models: Is it the Datasets or the Models?
Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane and Siva Reddy 14:30-16:00 (Regency A & B)
Knowledge-grounded conversational models are known to suffer from producing factually invalid statements, a phenomenon commonly called hallucination. In this work, we investigate the underlying causes of this phenomenon: is hallucination due to the training data, or to the models? We conduct a comprehensive human study on both existing knowledge-grounded conversational benchmarks and several state-of-the-art models. Our study reveals that the standard benchmarks consist of > 60% hallucinated responses, leading to models that not only hallucinate but even amplify hallucinations. Our findings raise important questions on the quality of existing datasets and models trained using them. We make our annotations publicly available for future research.

Show, Don't Tell: Demonstrations Outperform Descriptions for Schema-Guided Task-Oriented Dialogue
Raghav Gupta, Harrison Lee, Jeffrey Zhao, Yuan Cao, Abhinav Rastogi and Yonghui Wu 14:30-16:00 (Regency A & B)
Building universal dialogue systems that operate across multiple domains/APIs and generalize to new ones with minimal overhead is a critical challenge. Recent works have leveraged natural language descriptions of schema elements to enable such systems; however, descriptions only indirectly convey schema semantics. In this work, we propose Show, Don't Tell, which prompts seq2seq models with a labeled example dialogue to show the semantics of schema elements rather than tell the model through descriptions. While requiring similar effort from service developers as generating descriptions, we show that using short examples as schema representations with large language models results in state-of-the-art performance on two popular dialogue state tracking benchmarks designed to measure zero-shot generalization - the Schema-Guided Dialogue dataset and the MultiWOZ leave-one-out benchmark.

Generating Repetitions with Appropriate Repeated Words
Toshiki Kawamoto, Hidetaka Kamigaito, Kotaro Funakoshi and Manabu Okumura 14:30-16:00 (Regency A & B)
A repetition is a response that repeats words in the previous speaker's utterance in a dialogue. Repetitions are essential in communication to build trust with others, as investigated in linguistic studies. In this work, we focus on repetition generation. To the best of our knowledge, this is the first neural approach to address repetition generation. We propose Weighted Label Smoothing, a smoothing method for explicitly learning which words to repeat during fine-tuning, and a repetition scoring method that can output more appropriate repetitions during decoding. We conducted automatic and human evaluations involving applying these methods to the pre-trained language model T5 for generating repetitions. The experimental results indicate that our methods outperformed baselines in both evaluations.

ErAConD: Error Annotated Conversational Dialog Dataset for Grammatical Error Correction
Xun Yuan, Derek Pham, Sam Davidson and Zhou Yu 14:30-16:00 (Regency A & B)
Currently available grammatical error correction (GEC) datasets are compiled using essays or other long-form text written by language learners, limiting the applicability of these datasets to other domains such as informal writing and conversational dialog. In this paper, we present a novel GEC dataset consisting of parallel original and corrected utterances drawn from open-domain chatbot conversations; this dataset is, to our knowledge, the first GEC dataset targeted to a human-machine conversational setting. We also present a detailed annotation scheme which ranks errors by perceived impact on comprehension, making our dataset more representative of real-world language learning applications. To demonstrate the utility of the dataset, we use our annotated data to fine-tune a state-of-the-art GEC model. Experimental results show the effectiveness of our data in improving GEC model performance in a conversational scenario.

Incorporating Centering Theory into Neural Coreference Resolution
Haixia Chai and Michael Strube 14:30-16:00 (Regency A & B)
In recent years, transformer-based coreference resolution systems have achieved remarkable improvements on the CoNLL dataset. However, how coreference resolvers can benefit from discourse coherence is still an open question. In this paper, we propose to incorporate centering transitions derived from centering theory in the form of a graph into a neural coreference model. Our method improves the performance over the SOTA baselines, especially on pronoun resolution in long documents, formal well-structured text, and clusters with scattered mentions.

LEA: Meta Knowledge-Driven Self-Attentive Document Embedding for Few-Shot Text Classification
S. K. Hong and Tae Young Jang 14:30-16:00 (Regency A & B)
Text classification has achieved great success with the prosperity of deep learning and pre-trained language models. However, we often encounter labeled data deficiency problems in real-world text-classification tasks. To overcome such challenging scenarios, interest in few-shot learning has increased, whereas most few-shot text classification studies suffer from a difficulty of utilizing pre-trained language models. In the study, we propose a novel learning method for learning how to attend, called LEA, through which meta-level attention aspects are derived based on our meta-learning strategy. This enables the generation of task-specific document embedding with leveraging pre-trained language models even though a few labeled data instances are given. We evaluate our proposed learning method on five benchmark datasets. The results show that the novel method robustly provides the competitive performance compared to recent few-shot learning methods for all the datasets.

Event Schema Induction with Double Graph Autoencoders
Xiaomeng Jin, Manling Li and Heng Ji 14:30-16:00 (Regency A & B)
Event schema depicts the typical structure of complex events, serving as a scaffolding to effectively analyze, predict, and possibly intervene in the ongoing events. To induce event schemas from historical events, previous work uses an event-by-event scheme, ignoring the global structure of the entire schema graph. We propose a new event schema induction framework using double graph autoencoders, which captures the global dependencies among nodes in event graphs. Specifically, we first extract the event skeleton from an event graph and design a variational directed acyclic graph (DAG) autoencoder to learn its global structure. Then we further fill in the event arguments for the skeleton, and use another Graph Convolutional Network (GCN) based autoencoder to reconstruct entity-entity relations as well as to detect coreferential entities. By performing this two-stage induction decomposition, the model can avoid reconstructing the entire graph in one step, allowing it to focus on learning global structures between events. Experimental results on three event graph datasets demonstrate that our method achieves state-of-the-art performance and induces high-quality event schemas with global consistency.

Unified Semantic Typing with Meaningful Label Inference
James Y. Huang, Bangzheng Li, Jiashu Xu and Muhao Chen 14:30-16:00 (Regency A & B)

Semantic typing aims at classifying tokens or spans of interest in a textual context into semantic categories such as relations, entity types, and event types. The inferred labels of semantic categories meaningfully interpret how machines understand components of text. In this paper, we present UniST, a unified framework for semantic typing that captures label semantics by projecting both inputs and labels into a joint semantic embedding space. To formulate different lexical and relational semantic typing tasks as a unified task, we incorporate task descriptions to be jointly encoded with the input, allowing UniST to be adapted to different tasks without introducing task-specific model components. UniST optimizes a margin ranking loss such that the semantic relatedness of the input and labels is reflected from their embedding similarity. Our experiments demonstrate that UniST achieves strong performance across three semantic typing tasks: entity typing, relation classification and event typing. Meanwhile, UniST effectively transfers semantic knowledge of labels and substantially improves generalizability on inferring rarely seen and unseen types. In addition, multiple semantic typing tasks can be jointly trained within the unified framework, leading to a single compact multi-tasking model that performs comparably to dedicated single-task models, while offering even better transferability.

Crossroads, Buildings and Neighborhoods: A Dataset for Fine-grained Location Recognition
Pei Chen, Haotian Xu, Cheng Zhang and Ruihong Huang 14:30-16:00 (Regency A & B)
General domain Named Entity Recognition (NER) datasets like CoNLL-2003 mostly annotate coarse-grained location entities such as a country or a city. But many applications require identifying fine-grained locations from texts and mapping them precisely to geographic sites, e.g., a crossroad, an apartment building, or a grocery store. In this paper, we introduce a new dataset HarveyNER with fine-grained locations annotated in tweets. This dataset presents unique challenges and characterizes many complex and long location mentions in informal descriptions. We built strong baseline models using Curriculum Learning and experimented with different heuristic curricula to better recognize difficult location mentions. Experimental results show that the simple curricula can improve the system's performance on hard cases and its overall performance, and outperform several other baseline systems. The dataset and the baseline models can be found at https://github.com/brickee/HarveyNER.

CompactIE: Compact Facts in Open Information Extraction
Farima Fatahi Bayat, Nikita Bhutani and H. Jagadish 14:30-16:00 (Regency A & B)
A major drawback of modern neural OpenIE systems and benchmarks is that they prioritize high coverage of information in extractions over compactness of their constituents. This severely limits the usefulness of OpenIE extractions in many downstream tasks. The utility of extractions can be improved if extractions are compact and share constituents. To this end, we study the problem of identifying compact extractions with neural-based methods. We propose CompactIE, an OpenIE system that uses a novel pipelined approach to produce compact extractions with overlapping constituents. It first detects constituents of the extractions and then links them to build extractions. We train our system on compact extractions obtained by processing existing benchmarks. Our experiments on CaRB and Wire57 datasets indicate that CompactIE finds 1.5x-2x more compact extractions than previous systems, with high precision, establishing a new state-of-the-art performance in OpenIE.

Modeling Task Interactions in Document-Level Joint Entity and Relation Extraction
Liyan Xu and Jinho D. Choi 14:30-16:00 (Regency A & B)
We target on the document-level relation extraction in an end-to-end setting, where the model needs to jointly perform mention extraction, coreference resolution (COREF) and relation extraction (RE) at once, and gets evaluated in an entity-centric way. Especially, we address the two-way interaction between COREF and RE that has not been the focus by previous work, and propose to introduce explicit interaction namely Graph Compatibility (GC) that is specifically designed to leverage task characteristics, bridging decisions of two tasks for direct task interference. Our experiments are conducted on DocRED and DWIE; in addition to GC, we implement and compare different multi-task settings commonly adopted in previous work, including pipeline, shared encoders, graph propagation, to examine the effectiveness of different interactions. The result shows that GC achieves the best performance by up to 2.3/5.1 F1 improvement over the baseline.

Go Back in Time: Generating Flashbacks in Stories with Event Temporal Prompts
Rujun Han, Hong Chen, Yufei Tian and Nanyun Peng 14:30-16:00 (Regency A & B)
Stories or narratives are comprised of a sequence of events. To compose interesting stories, professional writers often leverage a creative writing technique called *flashback* that inserts past events into current storylines as we commonly observe in novels and plays. However, it is challenging for machines to generate *flashback* as it requires a solid understanding of event **temporal order** (e.g. *feeling hungry* before *eat*, not vice versa), and the creativity to arrange storylines so that earlier events do not always appear first in **narrative order**. Two major issues in existing systems that exacerbate the challenges: 1) temporal bias in pertaining and story datasets that leads to monotonic event temporal orders; 2) lack of explicit guidance that helps machines decide where to insert *flashbacks*. We propose to address these issues using structured storylines to encode events and their pair-wise temporal relations (before, after and vague) as **temporal prompts** that guide how stories should unfold temporally. We leverage a Plan-and-Write framework enhanced by reinforcement learning to generate storylines and stories end-to-end. Evaluation results show that the proposed method can generate more interesting stories with *flashbacks* while maintaining textual diversity, fluency, and temporal coherence.

Cross-Domain Detection of GPT-2-Generated Technical Text
Juan Diego Rodriguez, Todd Hay, David Gros, Zain Shamsi and Ravi Srinivasan 14:30-16:00 (Regency A & B)
Machine-generated text presents a potential threat not only to the public sphere, but also to the scientific enterprise, whereby genuine research is undermined by convincing, synthetic text. In this paper we examine the problem of detecting GPT-2-generated technical research text. We first consider the realistic scenario where the defender does not have full information about the adversary's text generation pipeline, but is able to label small amounts of in-domain genuine and synthetic text in order to adapt to the target distribution. Even in the extreme scenario of adapting a physics-domain detector to a biomedical detector, we find that only a few hundred labels are sufficient for good performance. Finally, we show that paragraph-level detectors can be used to detect the tampering of full-length documents under a variety of threat models.

Learning to Selectively Learn for Weakly Supervised Paraphrase Generation with Model-based Reinforcement Learning
Haiyan Yin, Dingcheng Li and Ping Li 14:30-16:00 (Regency A & B)
Paraphrase generation is an important language generation task attempting to interpret user intents and systematically generate new phrases of identical meanings to the given ones. However, the effectiveness of paraphrase generation is constrained by the access to the golden labeled data pairs where both the amount and the quality of the training data pairs are restricted. In this paper, we propose a new weakly supervised paraphrase generation approach that extends the success of a recent work that leverages reinforcement learning for effective model training with data selection. While data selection is privileged for the target task which has noisy data, developing a reinforced selective learning regime faces several unresolved challenges. In this paper, we carry on important discussions about the above problem and present a new model that could partially overcome the discussed issues with a model-based planning feature and a reward normalization feature. We perform extensive evaluation on four weakly supervised paraphrase generation tasks where the results show that our method could significantly improve the state-of-the-art performance on the evaluation datasets.

AmbiPun: Generating Humorous Puns with Ambiguous Context
Anirudh Mittal, Yufei Tian and Nanyun Peng 14:30-16:00 (Regency A & B)
In this paper, we propose a simple yet effective way to generate pun sentences that does not require any training on existing puns. Our

approach is inspired by humor theories that ambiguity comes from the context rather than the pun word itself. Given a pair of definitions of a pun word, our model first produces a list of related concepts through a reverse dictionary. We then utilize one-shot GPT3 to generate context words and then generate puns incorporating context words from both concepts. Human evaluation shows that our method successfully generates pun 52% of the time, outperforming well-crafted baselines and the state-of-the-art models by a large margin.

Inducing and Using Alignments for Transition-based AMR Parsing
Andrew Drozdov, Jiawei Zhou, Radu Florian, Andrew McCallum, Tahira Naseem, Yoon Kim and Ramón Fernandez Astudillo 14:30-16:00 (Regency A & B)
Transition-based parsers for Abstract Meaning Representation (AMR) rely on node-to-word alignments. These alignments are learned separately from parser training and require a complex pipeline of rule-based components, pre-processing, and post-processing to satisfy domain-specific constraints. Parsers also train on a point-estimate of the alignment pipeline, neglecting the uncertainty due to the inherent ambiguity of alignment. In this work we explore two avenues for overcoming these limitations. First, we propose a neural aligner for AMR that learns node-to-word alignments without relying on complex pipelines. We subsequently explore a tighter integration of aligner and parser training by considering a distribution over oracle action sequences arising from aligner uncertainty. Empirical results show this approach leads to more accurate alignments and generalization better from the AMR2.0 to AMR3.0 corpora. We attain a new state-of-the art for gold-only trained models, matching silver-trained performance without the need for beam search on AMR3.0.

Consistency Training with Virtual Adversarial Discrete Perturbation
Jungsoo Park, Gyuwan Kim and Jaewoo Kang 14:30-16:00 (Regency A & B)
Consistency training regularizes a model by enforcing predictions of original and perturbed inputs to be similar. Previous studies have proposed various augmentation methods for the perturbation but are limited in that they are agnostic to the training model. Thus, the perturbed samples may not aid in regularization due to their ease of classification from the model. In this context, we propose an augmentation method of adding a discrete noise that would incur the highest divergence between predictions. This virtual adversarial discrete noise obtained by replacing a small portion of tokens while keeping original semantics as much as possible efficiently pushes a training model's decision boundary. Experimental results show that our proposed method outperforms other consistency training baselines with text editing, paraphrasing, or a continuous noise on semi-supervised text classification tasks and a robustness benchmark.

Contrastive Learning for Prompt-based Few-shot Language Learners
Yiren Jian, Chongyang Gao and Soroush Vosoughi 14:30-16:00 (Regency A & B)
The impressive performance of GPT-3 using natural language prompts and in-context learning has inspired work on better fine-tuning of moderately-sized models under this paradigm. Following this line of work, we present a contrastive learning framework that clusters inputs from the same class for better generality of models trained with only limited examples. Specifically, we propose a supervised contrastive framework that clusters inputs from the same class under different augmented "views" and repel the ones from different classes. We create different "views" of an example by appending it with different language prompts and contextual demonstrations. Combining a contrastive loss with the standard masked language modeling (MLM) loss in prompt-based few-shot learners, the experimental results show that our method can improve over the state-of-the-art methods in a diverse set of 15 language tasks. Our framework makes minimal assumptions on the task or the base model, and can be applied to many recent methods with little modification.

Embedding Hallucination for Few-shot Language Fine-tuning
Yiren Jian, Chongyang Gao and Soroush Vosoughi 14:30-16:00 (Regency A & B)
Few-shot language learners adapt knowledge from a pre-trained model to recognize novel classes from a few-labeled sentences. In such settings, fine-tuning a pre-trained language model can cause severe over-fitting. In this paper, we propose an Embedding Hallucination (EmbedHalluc) method, which generates auxiliary embedding-label pairs to expand the fine-tuning dataset. The hallucinator is trained by playing an adversarial game with the discriminator, such that the hallucinated embedding is indiscriminative to the real ones in the fine-tuning dataset. By training with the extended dataset, the language learner effectively learns from the diverse hallucinated embeddings to overcome the over-fitting issue. Experiments demonstrate that our proposed method is effective in a wide range of language tasks, outperforming current fine-tuning methods. Further, we show that EmbedHalluc outperforms other methods that address this over-fitting problem, such as common data augmentation, semi-supervised pseudo-labeling, and regularization.

Exploring the Role of Task Transferability in Large-Scale Multi-Task Learning
Vishakh Padmakumar, Leonard Lausen, Miguel Ballesteros, Sheng Zha, He He and George Karypis 14:30-16:00 (Regency A & B)
Recent work has found that multi-task training with a large number of diverse tasks can uniformly improve downstream performance on unseen target tasks. In contrast, literature on task transferability has established that the choice of intermediate tasks can heavily affect downstream task performance. In this work, we aim to disentangle the effect of scale and relatedness of tasks in multi-task representation learning. We find that, on average, increasing the scale of multi-task learning, in terms of the number of tasks, indeed results in better learned representations than smaller multi-task setups. However, if the target tasks are known ahead of time, then training on a smaller set of related tasks is competitive to the large-scale multi-task training at a reduced computational cost.

Efficient Hierarchical Domain Adaptation for Pretrained Language Models
Alexandra Chronopoulou, Matthew E Peters and Jesse Dodge 14:30-16:00 (Regency A & B)
The remarkable success of large language models has been driven by dense models trained on massive unlabeled, unstructured corpora. These corpora typically contain text from diverse, heterogeneous sources, but information about the source of the text is rarely used during training. Transferring their knowledge to a target domain is typically done by continuing training in-domain. In this paper, we introduce a method to permit domain adaptation to many diverse domains using a computationally efficient adapter approach. Our method is based on the observation that textual domains are partially overlapping, and we represent domains as a hierarchical tree structure where each node in the tree is associated with a set of adapter weights. When combined with a frozen pretrained language model, this approach enables parameter sharing among related domains, while avoiding negative interference between unrelated ones. Experimental results with GPT-2 and a large fraction of the 100 most represented websites in C4 show across-the-board improvements in-domain. We additionally provide an inference time algorithm for a held-out domain and show that averaging over multiple paths through the tree enables further gains in generalization, while adding only a marginal cost to inference.

Learning Natural Language Generation with Truncated Reinforcement Learning
Alice Martin, Guillaume Quispe, Charles Ollion, Sylvain Le Corff, Florian Strub and Olivier Pietquin 14:30-16:00 (Regency A & B)
This paper introduces TRUncated ReinForcement Learning for Language (TrufLL), an original approach to train conditional language models without a supervised learning phase, by only using reinforcement learning (RL). As RL methods unsuccessfully scale to large action spaces, we dynamically truncate the vocabulary space using a generic language model. TrufLL thus enables to train a language agent by solely interacting with its environment without any task-specific prior knowledge; it is only guided with a task-agnostic language model. Interestingly, this approach avoids the dependency to labelled datasets and inherently reduces pretrained policy flaws such as language or exposure biases. We evaluate TrufLL on two visual question generation tasks, for which we report positive results over performance and language metrics,

which we then corroborate with a human evaluation. To our knowledge, it is the first approach that successfully learns a language generation policy without pre-training, using only reinforcement learning.

On the Effect of Pretraining Corpora on In-context Learning by a Large-scale Language Model
Seongjin Shin, Sang-Woo Lee, Hwijeen Ahn, Sungdong Kim, HyoungSeok Kim, Boseop Kim, Kyunghyun Cho, Gichang Lee, Woomyoung Park, Jung-Woo Ha and Nako Sung 14:30-16:00 (Regency A & B)
Many recent studies on large-scale language models have reported successful in-context zero- and few-shot learning ability. However, the in-depth analysis of when in-context learning occurs is still lacking. For example, it is unknown how in-context learning performance changes as the training corpus varies. Here, we investigate the effects of the source and size of the pretraining corpus on in-context learning in HyperCLOVA, a Korean-centric GPT-3 model. From our in-depth investigation, we introduce the following observations: (1) in-context learning performance heavily depends on the corpus domain source, and the size of the pretraining corpus does not necessarily determine the emergence of in-context learning, (2) in-context learning ability can emerge when a language model is trained on a combination of multiple corpora, even when each corpus does not result in in-context learning on its own, (3) pretraining with a corpus related to a downstream task does not always guarantee the competitive in-context learning performance of the downstream task, especially in the few-shot setting, and (4) the relationship between language modeling (measured in perplexity) and in-context learning does not always correlate: e.g., low perplexity does not always imply high in-context few-shot learning performance.

Learning to Generate Examples for Semantic Processing Tasks
Danilo Croce, Simone Filice, Giuseppe Castellucci and Roberto Basili 14:30-16:00 (Regency A & B)
Even if recent Transformer-based architectures, such as BERT, achieved impressive results in semantic processing tasks, their fine-tuning stage still requires large scale training resources. Usually, Data Augmentation (DA) techniques can help to deal with low resource settings. In Text Classification tasks, the objective of DA is the generation of well-formed sentences that i) represent the desired task category and ii) are novel with respect to existing sentences. In this paper, we propose a neural approach to automatically learn to generate new examples using a pre-trained sequence-to-sequence model. We first learn a task-oriented similarity function that we use to pair similar examples. Then, we use these example pairs to train a model to generate examples. Experiments in low resource settings show that augmenting the training material with the proposed strategy systematically improves the results on text classification and natural language inference tasks by up to 10% accuracy, outperforming existing DA approaches.

PARADISE: Exploiting Parallel Data for Multilingual Sequence-to-Sequence Pretraining
Machel Reid and Mikel Artetxe 14:30-16:00 (Regency A & B)
Despite the success of multilingual sequence-to-sequence pretraining, most existing approaches rely on monolingual corpora and do not make use of the strong cross-lingual signal contained in parallel data. In this paper, we present PARADISE (PARAllel &Denoising Integration in SEquence-to-sequence models), which extends the conventional denoising objective used to train these models by (i) replacing words in the noised sequence according to a multilingual dictionary, and (ii) predicting the reference translation according to a parallel corpus instead of recovering the original sequence. Our experiments on machine translation and cross-lingual natural language inference show an average improvement of 2.0 BLEU points and 6.7 accuracy points from integrating parallel data into pretraining, respectively, obtaining results that are competitive with several popular models at a fraction of their computational cost.

On Curriculum Learning for Commonsense Reasoning
Adyasha Maharana and Mohit Bansal 14:30-16:00 (Regency A & B)
Commonsense reasoning tasks follow a standard paradigm of finetuning pretrained language models on the target task data, where samples are introduced to the model in a random order during training. However, recent research suggests that data order can have a significant impact on the performance of finetuned models for natural language understanding. Hence, we examine the effect of a human-like easy-to-difficult curriculum during finetuning of language models for commonsense reasoning tasks. We use paced curriculum learning to rank data and sample training mini-batches with increasing levels of difficulty from the ranked dataset during finetuning. Further, we investigate the effect of an adaptive curriculum, i.e., the data ranking is dynamically updated during training based on the current state of the learner model. We use a teacher model to measure difficulty of each sample and experiment with three measures based on question answering probability, variability and out-of-distribution. To understand the effectiveness of curriculum learning in various scenarios, we apply it on full model fine-tuning as well as parameter-efficient prompt-tuning settings. Our results show that fixed as well as adaptive curriculum learning significantly improve performance for five commonsense reasoning tasks, i.e., SocialIQA, CosmosQA, CODAH, HellaSwag, WinoGrande in both tuning settings. Further, we find that prioritizing the difficult samples in the tail end of training improves generalization to unseen in-domain data as well as out-of-domain data. Our work provides evidence and encourages research into curriculum learning for commonsense reasoning.

Improving Multi-Document Summarization through Referenced Flexible Extraction with Credit-Awareness
Yun-Zhu Song, Yi-Syuan Chen and Hong-Han Shuai 14:30-16:00 (Regency A & B)
A notable challenge in Multi-Document Summarization (MDS) is the extremely-long length of the input. In this paper, we present an extract-then-abstract Transformer framework to overcome the problem. Specifically, we leverage pre-trained language models to construct a hierarchical extractor for salient sentence selection across documents and an abstractor for rewriting the selected contents as summaries. However, learning such a framework is challenging since the optimal contents for the abstractor are generally unknown. Previous works typically create *pseudo extraction oracle* to enable the supervised learning for both the extractor and the abstractor. Nevertheless, we argue that the performance of such methods could be restricted due to the insufficient information for prediction and inconsistent objectives between training and testing. To this end, we propose a loss weighting mechanism that makes the model aware of the unequal importance for the sentences not in the pseudo extraction oracle, and leverage the fine-tuned abstractor to generate summary references as auxiliary signals for learning the extractor. Moreover, we propose a reinforcement learning method that can efficiently apply to the extractor for harmonizing the optimization between training and testing. Experiment results show that our framework substantially outperforms strong baselines with comparable model sizes and achieves the best results on the Multi-News, Multi-XScience, and WikiCatSum corpora.

TSTR: Too Short to Represent, Summarize with Details! Intro-Guided Extended Summary Generation
Sajad Sotudeh and Nazli Goharian 14:30-16:00 (Regency A & B)
Many scientific papers such as those in arXiv and PubMed data collections have abstracts with varying lengths of 50-1000 words and average length of approximately 200 words, where longer abstracts typically convey more information about the source paper. Up to recently, scientific summarization research has typically focused on generating short, abstract-like summaries following the existing datasets used for scientific summarization. In domains where the source text is relatively long-form, such as in scientific documents, such summary is not able to go beyond the general and coarse overview and provide salient information from the source document. The recent interest to tackle this problem motivated curation of scientific datasets, arXiv-Long and PubMed-Long, containing human-written summaries of 400-600 words, hence, providing a venue for research in generating long/extended summaries. Extended summaries facilitate a faster read while providing details beyond coarse information. In this paper, we propose TSTR, an extractive summarizer that utilizes the introductory information of documents as pointers to their salient information. The evaluations on two existing large-scale extended summarization datasets indicate statistically significant improvement in terms of Rouge and average Rouge (F1) scores (except in one case) as compared to strong baselines

and state-of-the-art. Comprehensive human evaluations favor our generated extended summaries in terms of cohesion and completeness.

SueNes: A Weakly Supervised Approach to Evaluating Single-Document Summarization via Negative Sampling
Forrest Sheng Bao, Ge Luo, Hebi Li, Minghui Qiu, Yinfei Yang, Youbiao He and Cen Chen 14:30-16:00 (Regency A & B)
Canonical automatic summary evaluation metrics, such as ROUGE, focus on lexical similarity which cannot well capture semantics nor linguistic quality and require a reference summary which is costly to obtain. Recently, there have been a growing number of efforts to alleviate either or both of the two drawbacks. In this paper, we present a proof-of-concept study to a weakly supervised summary evaluation approach without the presence of reference summaries. Massive data in existing summarization datasets are transformed for training by pairing documents with corrupted reference summaries. In cross-domain tests, our strategy outperforms baselines with promising improvements, and show a great advantage in gauging linguistic qualities over all metrics.

Investigating Crowdsourcing Protocols for Evaluating the Factual Consistency of Summaries
Xiangru Tang, Alexander Fabbri, Haoran Li, Ziming Mao, Griffin Thomas Adams, Borui Wang, Asli Celikyilmaz, Yashar Mehdad and Dragomir Radev 14:30-16:00 (Regency A & B)
Current pre-trained models applied for summarization are prone to factual inconsistencies that misrepresent the source text. Evaluating the factual consistency of summaries is thus necessary to develop better models. However, the human evaluation setup for evaluating factual consistency has not been standardized. To determine the factors that affect the reliability of the human evaluation, we crowdsource evaluations for factual consistency across state-of-the-art models on two news summarization datasets using the rating-based Likert Scale and ranking-based Best-Worst Scaling. Our analysis reveals that the ranking-based Best-Worst Scaling offers a more reliable measure of summary quality across datasets and that the reliability of Likert ratings highly depends on the target dataset and the evaluation design. To improve crowdsourcing reliability, we extend the scale of the Likert rating and present a scoring algorithm for Best-Worst Scaling that we call value learning. Our crowdsourcing guidelines will be publicly available to facilitate future work on factual consistency in summarization.

Sort by Structure: Language Model Ranking as Dependency Probing
Max Müller-Eberstein, Rob Van Der Goot and Barbara Plank 14:30-16:00 (Regency A & B)
Making an informed choice of pre-trained language model (LM) is critical for performance, yet environmentally costly, and as such widely underexplored. The field of Computer Vision has begun to tackle encoder ranking, with promising forays into Natural Language Processing, however they lack coverage of linguistic tasks such as structured prediction. We propose probing to rank LMs, specifically for parsing dependencies in a given language, by measuring the degree to which labeled trees are recoverable from an LM's contextualized embeddings. Across 46 typologically and architecturally diverse LM-language pairs, our probing approach predicts the best LM choice 79% of the time using orders of magnitude less compute than training a full parser. Within this study, we identify and analyze one recently proposed decoupled LM—RemBERT—and find it strikingly contains less inherent dependency information, but often yields the best parser after full fine-tuning. Without this outlier our approach identifies the best LM in 89% of cases.

[CL] **The Impact of Edge Displacement Vaserstein Distance on UD Parsing Performance**
Mark Anderson and Carlos Gómez-Rodríguez 14:30-16:00 (Regency A & B)
We contribute to the discussion on parsing performance in NLP by introducing a measurement that evaluates the differences between the distributions of edge displacement (the directed distance of edges) seen in training and test data. We hypothesize that this measurement will be related to differences observed in parsing performance across treebanks. We motivate this by building upon previous work and then attempt to falsify this hypothesis by using a number of statistical methods. We establish that there is a statistical correlation between this measurement and parsing performance even when controlling for potential covariants. We then use this to establish a sampling technique that gives us an adversarial and complementary split. This gives an idea of the lower and upper bounds of parsing systems for a given treebank in lieu of freshly sampled data. In a broader sense, the methodology presented here can act as a reference for future correlation-based exploratory work in NLP.

Plenary Best Paper Awards & Land Acknowledgement

16:30-17:30 - **Auditorium** (Columbia C/D)

Main Conference: Tuesday, July 12, 2022

Session 3 - 08:00-09:00

Language Generation

08:00-09:00 (Columbia A)

Low Resource Style Transfer via Domain Adaptive Meta Learning
Xiangyang Li, Xiang Long, Yu Xia and Sujian Li 08:00-08:15 (Columbia A)
Text style transfer (TST) without parallel data has achieved some practical success. However, most of the existing unsupervised text style transfer methods suffer from (i) requiring massive amounts of non-parallel data to guide transferring different text styles. (ii) colossal performance degradation when fine-tuning the model in new domains. In this work, we propose DAML-ATM (Domain Adaptive Meta-Learning with Adversarial Transfer Model), which consists of two parts: DAML and ATM. DAML is a domain adaptive meta-learning approach to learn general knowledge in multiple heterogeneous source domains, capable of adapting to new unseen domains with a small amount of data. Moreover, we propose a new unsupervised TST approach Adversarial Transfer Model (ATM), composed of a sequence-to-sequence pre-trained language model and uses adversarial style training for better content preservation and style transfer. Results on multi-domain datasets demonstrate that our approach generalizes well on unseen low-resource domains, achieving state-of-the-art results against ten strong baselines.

Don't Take It Literally: An Edit-Invariant Sequence Loss for Text Generation
Guangyi Liu, Zichao Yang, Tianhua Tao, Xiaodan Liang, Junwei Bao, Zhen Li, Xiaodong He, Shuguang Cui and Zhiting Hu 08:15-08:30 (Columbia A)
Neural text generation models are typically trained by maximizing log-likelihood with the sequence cross entropy (CE) loss, which encourages an exact token-by-token match between a target sequence with a generated sequence. Such training objective is sub-optimal when the target sequence is not perfect, e.g., when the target sequence is corrupted with noises, or when only weak sequence supervision is available. To address the challenge, we propose a novel Edit-Invariant Sequence Loss (EISL), which computes the matching loss of a target n-gram with all n-grams in the generated sequence. EISL is designed to be robust to various noises and edits in the target sequences. Moreover, the EISL computation is essentially an approximate convolution operation with target n-grams as kernels, which is easy to implement and efficient to compute with existing libraries. To demonstrate the effectiveness of EISL, we conduct experiments on a wide range of tasks, including machine translation with noisy target sequences, unsupervised text style transfer with only weak training signals, and non-autoregressive generation with non-predefined generation order. Experimental results show our method significantly outperforms the common CE loss and other strong baselines on all the tasks. EISL has a simple API that can be used as a drop-in replacement of the CE loss: https://github.com/guangyliu/EISL.

An Exploration of Post-Editing Effectiveness in Text Summarization
Vivian Lai, Alison Smith-Renner, Ke Zhang, Ruijia Cheng, Wenjuan Zhang, Joel R. Tetreault and Alejandro Jaimes-Larrarte 08:30-08:45 (Columbia A)
Automatic summarization methods are efficient but can suffer from low quality. In comparison, manual summarization is expensive but produces higher quality. Can humans and AI collaborate to improve summarization performance? In similar text generation tasks (e.g., machine translation), human-AI collaboration in the form of "post-editing" AI-generated text reduces human workload and improves the quality of AI output. Therefore, we explored whether post-editing offers advantages in text summarization. Specifically, we conducted an experiment with 72 participants, comparing post-editing provided summaries with manual summarization for summary quality, human efficiency, and user experience on formal (XSum news) and informal (Reddit posts) text. This study sheds valuable insights on when post-editing is useful for text summarization: it helped in some cases (e.g., when participants lacked domain knowledge) but not in others (e.g., when provided summaries include inaccurate information). Participants' different editing strategies and needs for assistance offer implications for future human-AI summarization systems.

MOVER: Mask, Over-generate and Rank for Hyperbole Generation
Yunxiang Zhang and Xiaojun Wan 08:45-09:00 (Columbia A)
Despite being a common figure of speech, hyperbole is under-researched in Figurative Language Processing. In this paper, we tackle the challenging task of hyperbole generation to transfer a literal sentence into its hyperbolic paraphrase. To address the lack of available hyperbolic sentences, we construct HYPO-XL, the first large-scale English hyperbole corpus containing 17,862 hyperbolic sentences in a non-trivial way. Based on our corpus, we propose an unsupervised method for hyperbole generation that does not require parallel literal-hyperbole pairs. During training, we fine-tune BART to infill masked hyperbolic spans of sentences from HYPO-XL. During inference, we mask part of an input literal sentence and over-generate multiple possible hyperbolic versions. Then a BERT-based ranker selects the best candidate by hyperbolicity and paraphrase quality. Automatic and human evaluation results show that our model is effective at generating hyperbolic paraphrase sentences and outperforms several baseline systems.

Semantics & Sentiment Analysis

08:00-09:00 (Columbia C)

[TACL] **Fact Checking with Insufficient Evidence**
Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma and Isabelle Augenstein 08:00-08:15 (Columbia C)
Automating the fact checking (FC) process relies on information obtained from exter nal sources. In this work, we posit that it is crucial for FC models to make ve racity predictions only when there is suffi cient evidence and otherwise indicate when it is not enough. To this end, we are the first to study what information FC models consider sufficient by introducing a novel task and advancing it with three main con tributions. First, we conduct an in-depth empirical analysis of the task with a new fluency-preserving method for omitting in formation from the evidence at the con stituent and sentence level. We identify when models consider the remaining evi dence (in)sufficient for FC, based on three trained models with different Transformer architectures and three FC datasets. Sec ond, we ask annotators whether the omitted evidence was important for FC, resulting in a novel diagnostic dataset, SufficientFacts , for FC with omitted evidence. We find that models are least successful in detect ing missing evidence when adverbial mod ifiers are omitted (21% accuracy), whereas it is easiest for omitted

date modifiers (63% accuracy). Finally, we propose a novel data augmentation strategy for contrastive self- learning of missing evidence by employing the proposed omission method combined with tri-training. It improves performance for Evidence Sufficiency Prediction by up to 17.8 F 1 score, which in turn improves FC performance by up to 2.6 F 1 score.

A Double-Graph Based Framework for Frame Semantic Parsing
Ce Zheng, Xudong Chen, Runxin Xu and Baobao Chang 08:15-08:30 (Columbia C)
Frame semantic parsing is a fundamental NLP task, which consists of three subtasks: frame identification, argument identification and role classification. Most previous studies tend to neglect relations between different subtasks and arguments and pay little attention to ontological frame knowledge defined in FrameNet. In this paper, we propose a Knowledge-guided Incremental semantic parser with Double-graph (KID). We first introduce Frame Knowledge Graph (FKG), a heterogeneous graph containing both frames and FEs (Frame Elements) built on the frame knowledge so that we can derive knowledge-enhanced representations for frames and FEs. Besides, we propose Frame Semantic Graph (FSG) to represent frame semantic structures extracted from the text with graph structures. In this way, we can transform frame semantic parsing into an incremental graph construction problem to strengthen interactions between subtasks and relations between arguments. Our experiments show that KID outperforms the previous state-of-the-art method by up to 1.7 F1-score on two FrameNet datasets. Our code is availavle at https://github.com/PKUnlp-icler/KID.

Identifying Implicitly Abusive Remarks about Identity Groups using a Linguistically Informed Approach
Michael Wiegand, Elisabeth Eder and Josef Ruppenhofer 08:30-08:45 (Columbia C)
We address the task of distinguishing implicitly abusive sentences on identity groups ("Muslims contaminate our planet") from other group-related negative polar sentences ("Muslims despise terrorism"). Implicitly abusive language are utterances not conveyed by abusive words (e.g. "bimbo" or "scum"). So far, the detection of such utterances could not be properly addressed since existing datasets displaying a high degree of implicit abuse are fairly biased. Following the recently-proposed strategy to solve implicit abuse by separately addressing its different subtypes, we present a new focused and less biased dataset that consists of the subtype of atomic negative sentences about identity groups. For that task, we model components that each address one facet of such implicit abuse, i.e. depiction as perpetrators, aspectual classification and non-conformist views. The approach generalizes across different identity groups and languages.

Disentangled Learning of Stance and Aspect Topics for Vaccine Attitude Detection in Social Media
Lixing Zhu, Zheng Fang, Gabriele Pergola, Robert Procter and Yulan He 08:45-09:00 (Columbia C)
Building models to detect vaccine attitudes on social media is challenging because of the composite, often intricate aspects involved, and the limited availability of annotated data. Existing approaches have relied heavily on supervised training that requires abundant annotations and pre-defined aspect categories. Instead, with the aim of leveraging the large amount of unannotated data now available on vaccination, we propose a novel semi-supervised approach for vaccine attitude detection, called VADet. A variational autoencoding architecture based on language models is employed to learn from unlabelled data the topical information of the domain. Then, the model is fine-tuned with a few manually annotated examples of user attitudes. We validate the effectiveness of VADet on our annotated data and also on an existing vaccination corpus annotated with opinions on vaccines. Our results show that VADet is able to learn disentangled stance and aspect topics, and outperforms existing aspect-based sentiment analysis models on both stance detection and tweet clustering.

Language Resources & Evaluation 2

08:00-09:00 (Columbia D)

MuCGEC: a Multi-Reference Multi-Source Evaluation Dataset for Chinese Grammatical Error Correction
Yue Zhang, Zhenghua Li, Zuyi Bao, Jiacheng Li, Bo Zhang, Chen Li, Fei Huang and Min Zhang 08:00-08:15 (Columbia D)
This paper presents MuCGEC, a multi-reference multi-source evaluation dataset for Chinese Grammatical Error Correction (CGEC), consisting of 7,063 sentences collected from three Chinese-as-a-Second-Language (CSL) learner sources. Each sentence is corrected by three annotators, and their corrections are carefully reviewed by a senior annotator, resulting in 2.3 references per sentence. We conduct experiments with two mainstream CGEC models, i.e., the sequence-to-sequence model and the sequence-to-edit model, both enhanced with large pretrained language models, achieving competitive benchmark performance on previous and our datasets. We also discuss CGEC evaluation methodologies, including the effect of multiple references and using a char-based metric. Our annotation guidelines, data, and code are available at https://github.com/HillZhang1999/MuCGEC.

A Corpus for Understanding and Generating Moral Stories
Jian Guan, Ziqi Liu and Minlie Huang 08:15-08:30 (Columbia D)
Teaching morals is one of the most important purposes of storytelling. An essential ability for understanding and writing moral stories is bridging story plots and implied morals. Its challenges mainly lie in: (1) grasping knowledge about abstract concepts in morals, (2) capturing inter-event discourse relations in stories, and (3) aligning value preferences of stories and morals concerning good or bad behavior. In this paper, we propose two understanding tasks and two generation tasks to assess these abilities of machines. We present STORAL, a new dataset of Chinese and English human-written moral stories. We show the difficulty of the proposed tasks by testing various models with automatic and manual evaluation on STORAL. Furthermore, we present a retrieval-augmented algorithm that effectively exploits related concepts or events in training sets as additional guidance to improve performance on these tasks.

End-to-End Chinese Speaker Identification
Dian Yu, Ben Zhou and Dong Yu 08:30-08:45 (Columbia D)
Speaker identification (SI) in texts aims to identify the speaker(s) for each utterance in texts. Previous studies divide SI into several sub-tasks (e.g., quote extraction, named entity recognition, gender identification, and coreference resolution). However, we are still far from solving these sub-tasks, making SI systems that rely on them seriously suffer from error propagation. End-to-end SI systems, on the other hand, are not limited by individual modules, but suffer from insufficient training data from the existing small-scale datasets. To make large end-to-end models possible, we design a new annotation guideline that regards SI as span extraction from the local context, and we annotate by far the largest SI dataset for Chinese named CSI based on eighteen novels. Viewing SI as a span selection task also introduces the possibility of applying existing storng extractive machine reading comprehension (MRC) baselines. Surprisingly, simply using such a baseline without human-annotated character names and carefully designed rules, we can already achieve performance comparable or better than those of previous state-of-the-art SI methods on all public SI datasets for Chinese. Furthermore, we show that our dataset can serve as additional training data for existing benchmarks, which leads to further gains (up to 6.5% in accuracy). Finally, using CSI as a clean source, we design an effective self-training paradigm to continuously leverage hundreds of unlabeled novels.

WALNUT: A Benchmark on Semi-weakly Supervised Learning for Natural Language Understanding

Guoqing Zheng, Giannis Karamanolakis, Kai Shu and Ahmed Hassan Awadallah 08:45-09:00 (Columbia D)
Building machine learning models for natural language understanding (NLU) tasks relies heavily on labeled data. Weak supervision has been proven valuable when large amount of labeled data is unavailable or expensive to obtain. Existing works studying weak supervision for NLU either mostly focus on a specific task or simulate weak supervision signals from ground-truth labels. It is thus hard to compare different approaches and evaluate the benefit of weak supervision without access to a unified and systematic benchmark with diverse tasks and real-world weak labeling rules. In this paper, we propose such a benchmark, named WALNUT, to advocate and facilitate research on weak supervision for NLU. WALNUT consists of NLU tasks with different types, including document-level and token-level prediction tasks. WALNUT is the first semi-weakly supervised learning benchmark for NLU, where each task contains weak labels generated by multiple real-world weak sources, together with a small set of clean labels. We conduct baseline evaluations on WALNUT to systematically evaluate the effectiveness of various weak supervision methods and model architectures. Our results demonstrate the benefit of weak supervision for low-resource NLU tasks and highlight interesting patterns across tasks. We expect WALNUT to stimulate further research on methodologies to leverage weak supervision more effectively. The benchmark and code for baselines are available at aka.ms/walnut_benchmark.

Efficient Methods in NLP 2

08:00-09:00 (Elwha A)

Learning to Win Lottery Tickets in BERT Transfer via Task-agnostic Mask Training
Yuanxin Liu, Fandong Meng, Zheng Lin, Peng Fu, Yanan Cao, Weiping Wang and Jie Zhou 08:00-08:15 (Elwha A)
Recent studies on the lottery ticket hypothesis (LTH) show that pre-trained language models (PLMs) like BERT contain matching subnetworks that have similar transfer learning performance as the original PLM. These subnetworks are found using magnitude-based pruning. In this paper, we find that the BERT subnetworks have even more potential than these studies have shown. Firstly, we discover that the success of magnitude pruning can be attributed to the preserved pre-training performance, which correlates with the downstream transferability. Inspired by this, we propose to directly optimize the subnetwork structure towards the pre-training objectives, which can better preserve the pre-training performance. Specifically, we train binary masks over model weights on the pre-training tasks, with the aim of preserving the universal transferability of the subnetwork, which is agnostic to any specific downstream tasks. We then fine-tune the subnetworks on the GLUE benchmark and the SQuAD dataset. The results show that, compared with magnitude pruning, mask training can effectively find BERT subnetworks with improved overall performance on downstream tasks. Moreover, our method is also more efficient in searching subnetworks and more advantageous when fine-tuning within a certain range of data scarcity. Our code is available at https://github.com/llyx97/TAMT.

Knowledge Inheritance for Pre-trained Language Models
Yujia Qin, Yankai Lin, Jing Yi, Jiajie Zhang, Xu Han, Zhengyan Zhang, Yusheng Su, Zhiyuan Liu, Peng Li, Maosong Sun and Jie Zhou 08:15-08:30 (Elwha A)
Recent explorations of large-scale pre-trained language models (PLMs) have revealed the power of PLMs with huge amounts of parameters, setting off a wave of training ever-larger PLMs. However, it requires tremendous computational resources to train a large-scale PLM, which may be practically unaffordable. In addition, existing large-scale PLMs are mainly trained from scratch individually, ignoring that many well-trained PLMs are available. To this end, we explore the question how could existing PLMs benefit training large-scale PLMs in future. Specifically, we introduce a pre-training framework named "knowledge inheritance" (KI) and explore how could knowledge distillation serve as auxiliary supervision during pre-training to efficiently learn larger PLMs. Experimental results demonstrate the superiority of KI in training efficiency. We also conduct empirical analyses to explore the effects of teacher PLMs' pre-training settings, including model architecture, pre-training data, etc. Finally, we show that KI could be applied to domain adaptation and knowledge transfer.

Towards Efficient NLP: A Standard Evaluation and A Strong Baseline
Xiangyang Liu, Tianxiang Sun, Junliang He, Jiawen Wu, Lingling Wu, Xinyu Zhang, Hao Jiang, Zhao Cao, Xuanjing Huang and Xipeng Qiu 08:30-08:45 (Elwha A)
Supersized pre-trained language models have pushed the accuracy of various natural language processing (NLP) tasks to a new state-of-the-art (SOTA). Rather than pursuing the reachless SOTA accuracy, more and more researchers start paying attention to model efficiency and usability. Different from accuracy, the metric for efficiency varies across different studies, making them hard to be fairly compared. To that end, this work presents ELUE (Efficient Language Understanding Evaluation), a standard evaluation, and a public leaderboard for efficient NLP models. ELUE is dedicated to depicting the Pareto Frontier for various language understanding tasks, such that it can tell whether and how much a method achieves Pareto improvement. Along with the benchmark, we also release a strong baseline, ElasticBERT, which allows BERT to exit at any layer in both static and dynamic ways. We demonstrate the ElasticBERT, despite its simplicity, outperforms or performs on par with SOTA compressed and early exiting models. With ElasticBERT, the proposed ELUE has a strong Pareto Frontier and makes a better evaluation for efficient NLP models.

Adaptable Adapters
Nafise Sadat Moosavi, Quentin Delfosse, Kristian Kersting and Iryna Gurevych 08:45-09:00 (Elwha A)
State-of-the-art pretrained NLP models contain a hundred million to trillion parameters. Adapters provide a parameter-efficient alternative for the full finetuning in which we can only finetune lightweight neural network layers on top of pretrained weights. Adapter layers are initialized randomly. However, existing work uses the same adapter architecture—i.e., the same adapter layer on top of each layer of the pretrained model—for every dataset, regardless of the properties of the dataset or the amount of available training data. In this work, we introduce adaptable adapters that contain (1) learning different activation functions for different layers and different input data, and (2) a learnable switch to select and only use the beneficial adapter layers. We show that adaptable adapters achieve on-par performances with the standard adapter architecture while using a considerably smaller number of adapter layers. In addition, we show that the selected adapter architecture by adaptable adapters transfers well across different data settings and similar tasks. We propose to use adaptable adapters for designing efficient and effective adapter architectures. The resulting adapters (a) contain about 50% of the learning parameters of the standard adapter and are therefore more efficient at training and inference, and require less storage space, and (b) achieve considerably higher performances in low-data settings.

NLP Applications 2

08:00-09:00 (Elwha B)

CHEF: A Pilot Chinese Dataset for Evidence-Based Fact-Checking
Xuming Hu, Zhijiang Guo, GuanYu Wu, Aiwei Liu, Lijie Wen and Philip S. Yu 08:00-08:15 (Elwha B)
The explosion of misinformation spreading in the media ecosystem urges for automated fact-checking. While misinformation spans both geographic and linguistic boundaries, most work in the field has focused on English. Datasets and tools available in other languages, such as Chinese, are limited. In order to bridge this gap, we construct CHEF, the first CHinese Evidence-based Fact-checking dataset of 10K real-world claims. The dataset covers multiple domains, ranging from politics to public health, and provides annotated evidence retrieved from the Internet. Further, we develop established baselines and a novel approach that is able to model the evidence retrieval as a latent variable, allowing jointly training with the veracity prediction model in an end-to-end fashion. Extensive experiments show that CHEF will provide a challenging testbed for the development of fact-checking systems designed to retrieve and reason over non-English claims.

Progressive Class Semantic Matching for Semi-supervised Text Classification
Haiming Xu, Lingqiao Liu and Ehsan M Abbasnejad 08:15-08:30 (Elwha B)
Semi-supervised learning is a promising way to reduce the annotation cost for text-classification. Combining with pre-trained language models (PLMs), e.g., BERT, recent semi-supervised learning methods achieved impressive performance. In this work, we further investigate the marriage between semi-supervised learning and a pre-trained language model. Unlike existing approaches that utilize PLMs only for model parameter initialization, we explore the inherent topic matching capability inside PLMs for building a more powerful semi-supervised learning approach. Specifically, we propose a joint semi-supervised learning process that can progressively build a standard K-way classifier and a matching network for the input text and the Class Semantic Representation (CSR). The CSR will be initialized from the given labeled sentences and progressively updated through the training process. By means of extensive experiments, we show that our method can not only bring remarkable improvement to baselines, but also overall be more stable, and achieves state-of-the-art performance in semi-supervised text classification.

Unsupervised Paraphrasability Prediction for Compound Nominalizations
John Sie Yuen Lee, Ho Hung Lim and Carol Carol Webster 08:30-08:45 (Elwha B)
Commonly found in academic and formal texts, a nominalization uses a deverbal noun to describe an event associated with its corresponding verb. Nominalizations can be difficult to interpret because of ambiguous semantic relations between the deverbal noun and its arguments. Automatic generation of clausal paraphrases for nominalizations can help disambiguate their meaning. However, previous work has not identified cases where it is awkward or impossible to paraphrase a compound nominalization. This paper investigates unsupervised prediction of paraphrasability, which determines whether the prenominal modifier of a nominalization can be re-written as a noun or adverb in a clausal paraphrase. We adopt the approach of overgenerating candidate paraphrases followed by candidate ranking with a neural language model. In experiments on an English dataset, we show that features from an Abstract Meaning Representation graph lead to statistically significant improvement in both paraphrasability prediction and paraphrase generation.

Dual-Channel Evidence Fusion for Fact Verification over Texts and Tables
Nan Hu, Zirui Wu, Yuxuan Lai, Xiao Liu and Yansong Feng 08:45-09:00 (Elwha B)
Different from previous fact extraction and verification tasks that only consider evidence of a single format, FEVEROUS brings further challenges by extending the evidence format to both plain text and tables. Existing works convert all candidate evidence into either sentences or tables, thus often failing to fully capture the rich context in their original format from the converted evidence, let alone the context information lost during conversion. In this paper, we propose a Dual Channel Unified Format fact verification model (DCUF), which unifies various evidence into parallel streams, i.e., natural language sentences and a global evidence table, simultaneously. With carefully-designed evidence conversion and organization methods, DCUF makes the most of pre-trained table/language models to encourage each evidence piece to perform early and thorough interactions with other pieces in its original format. Experiments show that our model can make better use of existing pre-trained models to absorb evidence of two formats, thus outperforming previous works by a large margin. Our code and models are publicly available.

Session 4 - 10:45-12:15

Interpretability and Analysis of Models for NLP 1

10:45-12:15 (Columbia A)

Measure and Improve Robustness in NLP Models: A Survey
Xuezhi Wang, Haohan Wang and Diyi Yang 10:45-11:00 (Columbia A)
As NLP models achieved state-of-the-art performances over benchmarks and gained wide applications, it has been increasingly important to ensure the safe deployment of these models in the real world, e.g., making sure the models are robust against unseen or challenging scenarios. Despite robustness being an increasingly studied topic, it has been separately explored in applications like vision and NLP, with various definitions, evaluation and mitigation strategies in multiple lines of research. In this paper, we aim to provide a unifying survey of how to define, measure and improve robustness in NLP. We first connect multiple definitions of robustness, then unify various lines of work on identifying robustness failures and evaluating models' robustness. Correspondingly, we present mitigation strategies that are data-driven, model-driven, and inductive-prior-based, with a more systematic view of how to effectively improve robustness in NLP models. Finally, we conclude by outlining open challenges and future directions to motivate further research in this area.

Using Paraphrases to Study Properties of Contextual Embeddings
Laura Burdick, Jonathan K Kummerfeld and Rada Mihalcea 11:00-11:15 (Columbia A)
We use paraphrases as a unique source of data to analyze contextualized embeddings, with a particular focus on BERT. Because paraphrases naturally encode consistent word and phrase semantics, they provide a unique lens for investigating properties of embeddings. Using the Paraphrase Database's alignments, we study words within paraphrases as well as phrase representations. We find that contextual embeddings effectively handle polysemous words, but give synonyms surprisingly different representations in many cases. We confirm previous findings that BERT is sensitive to word order, but find slightly different patterns than prior work in terms of the level of contextualization across BERT's layers.

Can Rationalization Improve Robustness?
Howard Chen, Jacqueline He, Karthik R Narasimhan and Danqi Chen 11:15-11:30 (Columbia A)
A growing line of work has investigated the development of neural NLP models that can produce rationales–subsets of input that can explain their model predictions. In this paper, we ask whether such rationale models can provide robustness to adversarial attacks in addition to their

interpretable nature. Since these models need to first generate rationales ("rationalizer") before making predictions ("predictor"), they have the potential to ignore noise or adversarially added text by simply masking it out of the generated rationale. To this end, we systematically generate various types of 'AddText' attacks for both token and sentence-level rationalization tasks and perform an extensive empirical evaluation of state-of-the-art rationale models across five different tasks. Our experiments reveal that the rationale models promise to improve robustness over AddText attacks while they struggle in certain scenarios–when the rationalizer is sensitive to position bias or lexical choices of attack text. Further, leveraging human rationale as supervision does not always translate to better performance. Our study is a first step towards exploring the interplay between interpretability and robustness in the rationalize-then-predict framework.

Necessity and Sufficiency for Explaining Text Classifiers: A Case Study in Hate Speech Detection
Esma Balkir, Isar Nejadgholi, Kathleen C. Fraser and Svetlana Kiritchenko 11:30-11:45 (Columbia A)
We present a novel feature attribution method for explaining text classifiers, and analyze it in the context of hate speech detection. Although feature attribution models usually provide a single importance score for each token, we instead provide two complementary and theoretically-grounded scores – necessity and sufficiency – resulting in more informative explanations. We propose a transparent method that calculates these values by generating explicit perturbations of the input text, allowing the importance scores themselves to be explainable. We employ our method to explain the predictions of different hate speech detection models on the same set of curated examples from a test suite, and show that different values of necessity and sufficiency for identity terms correspond to different kinds of false positive errors, exposing sources of classifier bias against marginalized groups.

What do tokens know about their characters and how do they know it?
Ayush Kaushal and Kyle Mahowald 11:45-12:00 (Columbia A)
Pre-trained language models (PLMs) that use subword tokenization schemes can succeed at a variety of language tasks that require character-level information, despite lacking explicit access to the character composition of tokens. Here, studying a range of models (e.g., GPT- J, BERT, RoBERTa, GloVe), we probe what word pieces encode about character-level information by training classifiers to predict the presence or absence of a particular alphabetical character in a token, based on its embedding (e.g., probing whether the model embedding for "cat" encodes that it contains the character "a"). We find that these models robustly encode character-level information and, in general, larger models perform better at the task. We show that these results generalize to characters from non-Latin alphabets (Arabic, Devanagari, and Cyrillic). Then, through a series of experiments and analyses, we investigate the mechanisms through which PLMs acquire English-language character information during training and argue that this knowledge is acquired through multiple phenomena, including a systematic relationship between particular characters and particular parts of speech, as well as natural variability in the tokenization of related strings.

Do Prompt-Based Models Really Understand the Meaning of Their Prompts?
Albert Webson and Ellie Pavlick 12:00-12:15 (Columbia A)
Recently, a boom of papers has shown extraordinary progress in zero-shot and few-shot learning with various prompt-based models. It is commonly argued that prompts help models to learn faster in the same way that humans learn faster when provided with task instructions expressed in natural language. In this study, we experiment with over 30 prompts manually written for natural language inference (NLI). We find that models can learn just as fast with many prompts that are intentionally irrelevant or even pathologically misleading as they do with instructively "good" prompts. Further, such patterns hold even for models as large as 175 billion parameters (Brown et al., 2020) as well as the recently proposed instruction-tuned models which are trained on hundreds of prompts (Sanh et al., 2021). That is, instruction-tuned models often produce good predictions with irrelevant and misleading prompts even at zero shots. In sum, notwithstanding prompt-based models' impressive improvement, we find evidence of serious limitations that question the degree to which such improvement is derived from models understanding task instructions in ways analogous to humans' use of task instructions.

Summarization

10:45-12:15 (Columbia C)

NeuS: Neutral Multi-News Summarization for Mitigating Framing Bias
Nayeon Lee, Yejin Bang, Tiezheng YU, Andrea Madotto and Pascale Fung 10:45-11:00 (Columbia C)
Media news framing bias can increase political polarization and undermine civil society. The need for automatic mitigation methods is therefore growing. We propose a new task, a neutral summary generation from multiple news articles of the varying political leanings to facilitate balanced and unbiased news reading. In this paper, we first collect a new dataset, illustrate insights about framing bias through a case study, and propose a new effective metric and model (NeuS-Title) for the task. Based on our discovery that title provides a good signal for framing bias, we present NeuS-Title that learns to neutralize news content in hierarchical order from title to article. Our hierarchical multi-task learning is achieved by formatting our hierarchical data pair (title, article) sequentially with identifier-tokens ("TITLE=>", "ARTICLE=>") and fine-tuning the auto-regressive decoder with the standard negative log-likelihood objective. We then analyze and point out the remaining challenges and future directions. One of the most interesting observations is that neural NLG models can hallucinate not only factually inaccurate or unverifiable content but also politically biased content.

Joint Learning-based Heterogeneous Graph Attention Network for Timeline Summarization
Jingyi You, Dongyuan Li, Hidetaka Kamigaito, Kotaro Funakoshi and Manabu Okumura 11:00-11:15 (Columbia C)
Previous studies on the timeline summarization (TLS) task ignored the information interaction between sentences and dates, and adopted pre-defined unlearnable representations for them. They also considered date selection and event detection as two independent tasks, which makes it impossible to integrate their advantages and obtain a globally optimal summary. In this paper, we present a *joint learning-based heterogeneous graph attention network for TLS* (HeterTLS), in which date selection and event detection are combined into a unified framework to improve the extraction accuracy and remove redundant sentences simultaneously. Our heterogeneous graph involves multiple types of nodes, the representations of which are iteratively learned across the heterogeneous graph attention layer. We evaluated our model on four datasets, and found that it significantly outperformed the current state-of-the-art baselines with regard to ROUGE scores and date selection metrics.

Interactive Query-Assisted Summarization via Deep Reinforcement Learning
Ori Shapira, Ramakanth Pasunuru, Mohit Bansal, Ido Dagan and Yael Amsterdamer 11:15-11:30 (Columbia C)
Interactive summarization is a task that facilitates user-guided exploration of information within a document set. While one would like to employ state of the art neural models to improve the quality of interactive summarization, many such technologies cannot ingest the full document set or cannot operate at sufficient speed for interactivity. To that end, we propose two novel deep reinforcement learning models for the task that address, respectively, the subtask of summarizing salient information that adheres to user queries, and the subtask of listing suggested queries to assist users throughout their exploration. In particular, our models allow encoding the interactive session state and history

to refrain from redundancy. Together, these models compose a state of the art solution that addresses all of the task requirements. We compare our solution to a recent interactive summarization system, and show through an experimental study involving real users that our models are able to improve informativeness while preserving positive user experience.

FactPEGASUS: Factuality-Aware Pre-training and Fine-tuning for Abstractive Summarization
David Wan and Mohit Bansal 11:30-11:45 (Columbia C)
We present FactPEGASUS, an abstractive summarization model that addresses the problem of factuality during pre-training and fine-tuning: (1) We augment the sentence selection strategy of PEGASUS's (Zhang et al., 2019) pre-training objective to create pseudo-summaries that are both important and factual; (2) We introduce three complementary components for fine-tuning. The corrector removes hallucinations present in the reference summary, the contrastor uses contrastive learning to better differentiate nonfactual summaries from factual ones, and the connector bridges the gap between the pre-training and fine-tuning for better transfer of knowledge. Experiments on three downstream tasks demonstrate that FactPEGASUS substantially improves factuality evaluated by multiple automatic metrics and humans. Our thorough analysis suggests that FactPEGASUS is more factual than using the original pre-training objective in zero-shot and few-shot settings, retains factual behavior more robustly than strong baselines, and does not rely entirely on becoming more extractive to improve factuality.

Proposition-Level Clustering for Multi-Document Summarization
Ori Ernst, Avi Caciularu, Ori Shapira, Ramakanth Pasunuru, Mohit Bansal, Jacob Goldberger and Ido Dagan 11:45-12:00 (Columbia C)
Text clustering methods were traditionally incorporated into multi-document summarization (MDS) as a means for coping with considerable information repetition. Particularly, clusters were leveraged to indicate information saliency as well as to avoid redundancy. Such prior methods focused on clustering sentences, even though closely related sentences usually contain also non-aligned parts. In this work, we revisit the clustering approach, grouping together sub-sentential propositions, aiming at more precise information alignment. Specifically, our method detects salient propositions, clusters them into paraphrastic clusters, and generates a representative sentence for each cluster via text fusion. Our summarization method improves over the previous state-of-the-art MDS method in the DUC 2004 and TAC 2011 datasets, both in automatic ROUGE scores and human preference.

[TACL] **A Multi-Level Optimization Framework for End-to-End Text Augmentation**
Sai Ashish Somayajula, Linfeng Song and Pengtao Xie 12:00-12:15 (Columbia C)
Text augmentation is an effective technique in alleviating overfitting in NLP tasks. In existing methods, text augmentation and downstream tasks are mostly performed separately. As a result, the augmented texts may not be optimal to train the downstream model. To address this problem, we propose a three-level optimization framework to perform text augmentation and the downstream task end-to-end. The augmentation model is trained in a way tailored to the downstream task. Our framework consists of three learning stages. A text summarization model is trained to perform data augmentation at the first stage. Each summarization example is associated with a weight to account for its domain difference with the text classification data. At the second stage, we use the model trained at the first stage to perform text augmentation and train a text classification model on the augmented texts. At the third stage, we evaluate the text classification model trained at the second stage and update weights of summarization examples by minimizing the validation loss. These three stages are performed end-to-end. We evaluate our method on several text classification datasets where the results demonstrate the effectiveness of our method.

Information Retrieval

10:45-12:15 (Columbia D)

CERES: Pretraining of Graph-Conditioned Transformer for Semi-Structured Session Data
Rui Feng, Chen Luo, Qingyu Yin, Bing Yin, Tuo Zhao and Chao Zhang 10:45-11:00 (Columbia D)
User sessions empower many search and recommendation tasks on a daily basis. Such session data are semi-structured, which encode heterogeneous relations between queries and products, and each item is described by the unstructured text. Despite recent advances in self-supervised learning for text or graphs, there lack of self-supervised learning models that can effectively capture both intra-item semantics and inter-item interactions for semi-structured sessions. To fill this gap, we propose CERES, a graph-based transformer model for semi-structured session data. CERES learns representations that capture both inter- and intra-item semantics with (1) a graph-conditioned masked language pretraining task that jointly learns from item text and item-item relations; and (2) a graph-conditioned transformer architecture that propagates inter-item contexts to item-level representations. We pretrained CERES using 468 million Amazon sessions and find that CERES outperforms strong pretraining baselines by up to 9% in three session search and entity linking tasks.

GPL: Generative Pseudo Labeling for Unsupervised Domain Adaptation of Dense Retrieval
Kexin Wang, Nandan Thakur, Nils Reimers and Iryna Gurevych 11:00-11:15 (Columbia D)
Dense retrieval approaches can overcome the lexical gap and lead to significantly improved search results. However, they require large amounts of training data which is not available for most domains. As shown in previous work (Thakur et al., 2021b), the performance of dense retrievers severely degrades under a domain shift. This limits the usage of dense retrieval approaches to only a few domains with large training datasets. In this paper, we propose the novel unsupervised domain adaptation method *Generative Pseudo Labeling* (GPL), which combines a query generator with pseudo labeling from a cross-encoder. On six representative domain-specialized datasets, we find the proposed GPL can outperform an out-of-the-box state-of-the-art dense retrieval approach by up to 9.3 points nDCG@10. GPL requires less (unlabeled) data from the target domain and is more robust in its training than previous methods. We further investigate the role of six recent pre-training methods in the scenario of domain adaptation for retrieval tasks, where only three could yield improved results. The best approach, TSDAE (Wang et al., 2021) can be combined with GPL, yielding another average improvement of 1.4 points nDCG@10 across the six tasks. The code and the models are available at https://github.com/UKPLab/gpl.

Boosted Dense Retriever
Patrick Lewis, Barlas Oguz, Wenhan Xiong, Fabio Petroni, Scott Yih and Sebastian Riedel 11:15-11:30 (Columbia D)
We propose DrBoost, a dense retrieval ensemble inspired by boosting. DrBoost is trained in stages: each component model is learned sequentially and specialized by focusing only on retrieval mistakes made by the current ensemble. The final representation is the concatenation of the output vectors of all the component models, making it a drop-in replacement for standard dense retrievers at test time. DrBoost enjoys several advantages compared to standard dense retrieval models. It produces representations which are 4x more compact, while delivering comparable retrieval results. It also performs surprisingly well under approximate search with coarse quantization, reducing latency and bandwidth needs by another 4x. In practice, this can make the difference between serving indices from disk versus from memory, paving the way for much cheaper deployments.

A Dataset for N-ary Relation Extraction of Drug Combinations

Aryeh Tiktinsky, Vijay Viswanathan, Danna Niezni, Dana Meron Azagury, Yosi Shamay, Hillel Taub-Tabib, Tom Hope and Yoav Goldberg
11:30-11:45 (Columbia D)
Combination therapies have become the standard of care for diseases such as cancer, tuberculosis, malaria and HIV. However, the combinatorial set of available multi-drug treatments creates a challenge in identifying effective combination therapies available in a situation. To assist medical professionals in identifying beneficial drug-combinations, we construct an expert-annotated dataset for extracting information about the efficacy of drug combinations from the scientific literature. Beyond its practical utility, the dataset also presents a unique NLP challenge, as the first relation extraction dataset consisting of variable-length relations. Furthermore, the relations in this dataset predominantly require language understanding beyond the sentence level, adding to the challenge of this task. We provide a promising baseline model and identify clear areas for further improvement. We release our dataset (https://huggingface.co/datasets/allenai/drug-combo-extraction), code (https://github.com/allenai/drug-combo-extraction) and baseline models (https://huggingface.co/allenai/drug-combo-classifier-pubmedbert-dapt) publicly to encourage the NLP community to participate in this task.

ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction
Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts and Matei Zaharia 11:45-12:00 (Columbia D)
Neural information retrieval (IR) has greatly advanced search and other knowledge-intensive language tasks. While many neural IR methods encode queries and documents into single-vector representations, late interaction models produce multi-vector representations at the granularity of each token and decompose relevance modeling into scalable token-level computations. This decomposition has been shown to make late interaction more effective, but it inflates the space footprint of these models by an order of magnitude. In this work, we introduce Maize, a retriever that couples an aggressive residual compression mechanism with a denoised supervision strategy to simultaneously improve the quality and space footprint of late interaction. We evaluate Maize across a wide range of benchmarks, establishing state-of-the-art quality within and outside the training domain while reducing the space footprint of late interaction models by 6–10x.

Multi-Vector Models with Textual Guidance for Fine-Grained Scientific Document Similarity
Sheshera Mysore, Arman Cohan and Tom Hope 12:00-12:15 (Columbia D)
We present a new scientific document similarity model based on matching fine-grained aspects of texts. To train our model, we exploit a naturally-occurring source of supervision: sentences in the full-text of papers that cite multiple papers together (co-citations). Such co-citations not only reflect close paper relatedness, but also provide textual descriptions of how the co-cited papers are related. This novel form of textual supervision is used for learning to match aspects across papers. We develop multi-vector representations where vectors correspond to sentence-level aspects of documents, and present two methods for aspect matching: (1) A fast method that only matches single aspects, and (2) a method that makes sparse multiple matches with an Optimal Transport mechanism that computes an Earth Mover's Distance between aspects. Our approach improves performance on document similarity tasks in four datasets. Further, our fast single-match method achieves competitive results, paving the way for applying fine-grained similarity to large scientific corpora.

Language Grounding to Vision 1

10:45-12:15 (Elwha A)

KAT: A Knowledge Augmented Transformer for Vision-and-Language
Liangke Gui, Borui Wang, Qiuyuan Huang, Alexander G Hauptmann, Yonatan Bisk and Jianfeng Gao 10:45-11:00 (Elwha A)
The primary focus of recent work with large-scale transformers has been on optimizing the amount of information packed into the model's parameters. In this work, we ask a complementary question: Can multimodal transformers leverage explicit knowledge in their reasoning? Existing, primarily unimodal, methods have explored approaches under the paradigm of knowledge retrieval followed by answer prediction, but leave open questions about the quality and relevance of the retrieved knowledge used, and how the reasoning processes over implicit and explicit knowledge should be integrated. To address these challenges, we propose a - Knowledge Augmented Transformer (KAT) - which achieves a strong state-of-the-art result (+6% absolute) on the open-domain multimodal task of OK-VQA. Our approach integrates implicit and explicit knowledge in an encoder-decoder architecture, while still jointly reasoning over both knowledge sources during answer generation. Additionally, explicit knowledge integration improves interpretability of model predictions in our analysis.

Do Trajectories Encode Verb Meaning?
Dylan Ebert, Chen Sun and Ellie Pavlick 11:00-11:15 (Elwha A)
Distributional models learn representations of words from text, but are criticized for their lack of grounding, or the linking of text to the non-linguistic world. Grounded language models have had success in learning to connect concrete categories like nouns and adjectives to the world via images and videos, but can struggle to isolate the meaning of the verbs themselves from the context in which they typically occur. In this paper, we investigate the extent to which trajectories (i.e. the position and rotation of objects over time) naturally encode verb semantics. We build a procedurally generated agent-object-interaction dataset, obtain human annotations for the verbs that occur in this data, and compare several methods for representation learning given the trajectories. We find that trajectories correlate as-is with some verbs (e.g., fall), and that additional abstraction via self-supervised pretraining can further capture nuanced differences in verb meaning (e.g., roll and slide).

Diagnosing Vision-and-Language Navigation: What Really Matters
Wanrong Zhu, Yuankai Qi, Pradyumna Narayana, Kazoo Sone, Sugato Basu, Xin Eric Wang, Qi Wu, Miguel Eckstein and William Yang Wang
11:15-11:30 (Elwha A)
Vision-and-language navigation (VLN) is a multimodal task where an agent follows natural language instructions and navigates in visual environments. Multiple setups have been proposed, and researchers apply new model architectures or training techniques to boost navigation performance. However, there still exist non-negligible gaps between machines' performance and human benchmarks. Moreover, the agents' inner mechanisms for navigation decisions remain unclear. To the best of our knowledge, how the agents perceive the multimodal input is under-studied and needs investigation. In this work, we conduct a series of diagnostic experiments to unveil agents' focus during navigation. Results show that indoor navigation agents refer to both object and direction tokens when making decisions. In contrast, outdoor navigation agents heavily rely on direction tokens and poorly understand the object tokens. Transformer-based agents acquire a better cross-modal understanding of objects and display strong numerical reasoning ability than non-Transformer-based agents. When it comes to vision-and-language alignments, many models claim that they can align object tokens with specific visual targets. We find unbalanced attention on the vision and text input and doubt the reliability of such cross-modal alignments.

Connecting the Dots between Audio and Text without Parallel Data through Visual Knowledge Transfer
Yanpeng Zhao, Jack Hessel, Youngjae Yu, Ximing Lu, Rowan Zellers and Yejin Choi 11:30-11:45 (Elwha A)
Machines that can represent and describe environmental soundscapes have practical potential, e.g., for audio tagging and captioning. Prevail-

ing learning paradigms of audio-text connections have been relying on parallel audio-text data, which is, however, scarcely available on the web. We propose VIP-ANT that induces Audio-Text alignment without using any parallel audio-text data. Our key idea is to share the image modality between bi-modal image-text representations and bi-modal image-audio representations; the image modality functions as a pivot and connects audio and text in a tri-modal embedding space implicitly. In a difficult zero-shot setting with no paired audio-text data, our model demonstrates state-of-the-art zero-shot performance on the ESC50 and US8K audio classification tasks, and even surpasses the supervised state of the art for Clotho caption retrieval (with audio queries) by 2.2% R@1. We further investigate cases of minimal audio-text supervision, finding that, e.g., just a few hundred supervised audio-text pairs increase the zero-shot audio classification accuracy by 8% on US8K. However, to match human parity on some zero-shot tasks, our empirical scaling experiments suggest that we would need about $2^{21} \approx 2\mathrm{M}$ supervised audio-caption pairs. Our work opens up new avenues for learning audio-text connections with little to no parallel audio-text data.

Guiding Visual Question Generation
Nihir Vedd, Zixu Wang, Marek Rei, Yishu Miao and Lucia Specia 11:45-12:00 (Elwha A)
In traditional Visual Question Generation (VQG), most images have multiple concepts (e.g. objects and categories) for which a question could be generated, but models are trained to mimic an arbitrary choice of concept as given in their training data. This makes training difficult and also poses issues for evaluation – multiple valid questions exist for most images but only one or a few are captured by the human references. We present Guiding Visual Question Generation - a variant of VQG which conditions the question generator on categorical information based on expectations on the type of question and the objects it should explore. We propose two variant families: (i) an explicitly guided model that enables an actor (human or automated) to select which objects and categories to generate a question for; and (ii) 2 types of implicitly guided models that learn which objects and categories to condition on, based on discrete variables. The proposed models are evaluated on an answer-category augmented VQA dataset and our quantitative results show a substantial improvement over the current state of the art (over 9 BLEU-4 increase). Human evaluation validates that guidance helps the generation of questions that are grammatically coherent and relevant to the given image and objects.

Interactive Symbol Grounding with Complex Referential Expressions
Rimvydas Rubavicius and Alex Lascarides 12:00-12:15 (Elwha A)
We present a procedure for learning to ground symbols from a sequence of stimuli consisting of an arbitrarily complex noun phrase (e.g. “all but one green square above both red circles.”) and its designation in the visual scene. Our distinctive approach combines: a) lazy few-shot learning to relate open-class words like `green` and `above` to their visual percepts; and b) symbolic reasoning with closed-class word categories like quantifiers and negation. We use this combination to estimate new training examples for grounding symbols that occur *within* a noun phrase but aren’t designated by that noun phase (e.g, `red` in the above example), thereby potentially gaining data efficiency. We evaluate the approach in a visual reference resolution task, in which the learner starts out unaware of concepts that are part of the domain model and how they relate to visual percepts.

Dialogue and Interactive Systems 1

10:45-12:15 (Elwha B)

Learning to Express in Knowledge-Grounded Conversation
Xueliang Zhao, Tingchen Fu, Chongyang Tao, Wei Wu, Dongyan Zhao and Rui Yan 10:45-11:00 (Elwha B)
Grounding dialogue generation by extra knowledge has shown great potentials towards building a system capable of replying with knowledgeable and engaging responses. Existing studies focus on how to synthesize a response with proper knowledge, yet neglect that the same knowledge could be expressed differently by speakers even under the same context. In this work, we mainly consider two aspects of knowledge expression, namely the structure of the response and style of the content in each part. We therefore introduce two sequential latent variables to represent the structure and the content style respectively. We propose a segmentation-based generation model and optimize the model by a variational approach to discover the underlying pattern of knowledge expression in a response. Evaluation results on two benchmarks indicate that our model can learn the structure style defined by a few examples and generate responses in desired content style.

Partner Personas Generation for Dialogue Response Generation
Hongyuan Lu, Wai Lam, Hong Cheng and Helen M. Meng 11:00-11:15 (Elwha B)
Incorporating personas information allows diverse and engaging responses in dialogue response generation. Unfortunately, prior works have primarily focused on self personas and have overlooked the value of partner personas. Moreover, in practical applications, the availability of the gold partner personas is often not the case. This paper attempts to tackle these issues by offering a novel framework that leverages automatic partner personas generation to enhance the succeeding dialogue response generation. Our framework employs reinforcement learning with a dedicatedly designed critic network for reward judgement. Experimental results from automatic and human evaluations indicate that our framework is capable of generating relevant, interesting, coherent and informative partner personas, even compared to the ground truth partner personas. This enhances the succeeding dialogue response generation, which surpasses our competitive baselines that condition on the ground truth partner personas.

Robust Conversational Agents against Imperceptible Toxicity Triggers
Ninareh Mehrabi, Ahmad Beirami, Fred Morstatter and Aram Galstyan 11:15-11:30 (Elwha B)
Warning: this paper contains content that maybe offensive or upsetting. Recent research in Natural Language Processing (NLP) has advanced the development of various toxicity detection models with the intention of identifying and mitigating toxic language from existing systems. Despite the abundance of research in this area, less attention has been given to adversarial attacks that force the system to generate toxic language and the defense against them. Existing work to generate such attacks is either based on human-generated attacks which is costly and not scalable or, in case of automatic attacks, the attack vector does not conform to human-like language, which can be detected using a language model loss. In this work, we propose attacks against conversational agents that are imperceptible, i.e., they fit the conversation in terms of coherency, relevancy, and fluency, while they are effective and scalable, i.e., they can automatically trigger the system into generating toxic language. We then propose a defense mechanism against such attacks which not only mitigates the attack but also attempts to maintain the conversational flow. Through automatic and human evaluations, we show that our defense is effective at avoiding toxic language generation even against imperceptible toxicity triggers while the generated language fits the conversation in terms of coherency and relevancy. Lastly, we establish the generalizability of such a defense mechanism on language generation models beyond conversational agents.

VGNMN: Video-grounded Neural Module Networks for Video-Grounded Dialogue Systems
Hung Le, Nancy F. Chen and Steven Hoi 11:30-11:45 (Elwha B)
Neural module networks (NMN) have achieved success in image-grounded tasks such as Visual Question Answering (VQA) on synthetic images. However, very limited work on NMN has been studied in the video-grounded dialogue tasks. These tasks extend the complexity

of traditional visual tasks with the additional visual temporal variance and language cross-turn dependencies. Motivated by recent NMN approaches on image-grounded tasks, we introduce Video-grounded Neural Module Network (VGNMN) to model the information retrieval process in video-grounded language tasks as a pipeline of neural modules. VGNMN first decomposes all language components in dialogues to explicitly resolve any entity references and detect corresponding action-based inputs from the question. The detected entities and actions are used as parameters to instantiate neural module networks and extract visual cues from the video. Our experiments show that VGNMN can achieve promising performance on a challenging video-grounded dialogue benchmark as well as a video QA benchmark.

Multimodal Dialogue State Tracking
Hung Le, Nancy F. Chen and Steven Hoi 11:45-12:00 (Elwha B)
Designed for tracking user goals in dialogues, a dialogue state tracker is an essential component in a dialogue system. However, the research of dialogue state tracking has largely been limited to unimodality, in which slots and slot values are limited by knowledge domains (e.g. restaurant domain with slots of restaurant name and price range) and are defined by specific database schema. In this paper, we propose to extend the definition of dialogue state tracking to multimodality. Specifically, we introduce a novel dialogue state tracking task to track the information of visual objects that are mentioned in video-grounded dialogues. Each new dialogue utterance may introduce a new video segment, new visual objects, or new object attributes and a state tracker is required to update these information slots accordingly. We created a new synthetic benchmark and designed a novel baseline, Video-Dialogue Transformer Network (VDTN), for this task. VDTN combines both object-level features and segment-level features and learns contextual dependencies between videos and dialogues to generate multimodal dialogue states. We optimized VDTN for a state generation task as well as a self-supervised video understanding task which recovers video segment or object representations. Finally, we trained VDTN to use the decoded states in a response prediction task. Together with comprehensive ablation and qualitative analysis, we discovered interesting insights towards building more capable multimodal dialogue systems.

Commonsense and Named Entity Aware Knowledge Grounded Dialogue Generation
Deeksha Varshney, Akshara Prabhakar and Asif Ekbal 12:00-12:15 (Elwha B)
Grounding dialogue on external knowledge and interpreting linguistic patterns in dialogue history context, such as ellipsis, anaphora, and co-reference is critical for dialogue comprehension and generation. In this paper, we present a novel open-domain dialogue generation model which effectively utilizes the large-scale commonsense and named entity based knowledge in addition to the unstructured topic-specific knowledge associated with each utterance. We enhance the commonsense knowledge with named entity-aware structures using co-references. Our proposed model utilizes a multi-hop attention layer to preserve the most accurate and critical parts of the dialogue history and the associated knowledge. In addition, we employ a Commonsense and Named Entity Enhanced Attention Module, which starts with the extracted triples from various sources and gradually finds the relevant supporting set of triples using multi-hop attention with the query vector obtained from the interactive dialogue-knowledge module. Empirical results on two benchmark datasets demonstrate that our model significantly outperforms the state-of-the-art methods in terms of both automatic evaluation metrics and human judgment. Our code is publicly available at https://github.com/deekshaVarshney/CNTF; https://www.iitp.ac.in/-ai-nlp-ml/resources/codes/CNTF.zip.

Industry Oral Session

08:00-09:00 (Quinault)

CREATER: CTR-driven Advertising Text Generation with Controlled Pre-Training and Contrastive Fine-Tuning
Penghui Wei, Xuanhua Yang, ShaoGuo Liu, Liang Wang and Bo Zheng 08:00-08:15 (Quinault)
This paper focuses on automatically generating the text of an ad, and the goal is that the generated text can capture user interest for achieving higher click-through rate (CTR). We propose CREATER, a CTR-driven advertising text generation approach, to generate ad texts based on high-quality user reviews. To incorporate CTR objective, our model learns from online A/B test data with contrastive learning, which encourages the model to generate ad texts that obtain higher CTR. To make use of large-scale unpaired reviews, we design a customized self-supervised objective reducing the gap between pre-training and fine-tuning. Experiments on industrial datasets show that CREATER significantly outperforms current approaches. It has been deployed online in a leading advertising platform and brings uplift on core online metrics.

Augmenting Poetry Composition with Verse by Verse
David Uthus, Maria Voitovich and R.J. Mical 08:30-08:15 (Quinault)
We describe Verse by Verse, our experiment in augmenting the creative process of writing poetry with an AI. We have created a group of AI poets, styled after various American classic poets, that are able to offer as suggestions generated lines of verse while a user is composing a poem. In this paper, we describe the underlying system to offer these suggestions. This includes a generative model, which is tasked with generating a large corpus of lines of verse offline and which are then stored in an index, and a dual-encoder model that is tasked with recommending the next possible set of verses from our index given the previous line of verse.

FPI: Failure Point Isolation in Large-scale Conversational Assistants
Rinat Khaziev, Usman Shahid, Tobias Röding, Rakesh Chada, Emir Kapanci and Pradeep Natarajan 08:30-08:45 (Quinault)
Large-scale conversational assistants such as Cortana, Alexa, Google Assistant and Siri process requests through a series of modules for wake word detection, speech recognition, language understanding and response generation. An error in one of these modules can cascade through the system. Given the large traffic volumes in these assistants, it is infeasible to manually analyze the data, identify requests with processing errors and isolate the source of error. We present a machine learning system to address this challenge. First, we embed the incoming request and context, such as system response and subsequent turns, using pre-trained transformer models. Then, we combine these embeddings with encodings of additional metadata features (such as confidence scores from different modules in the online system) using a "mixing-encoder" to output the failure point predictions. Our system obtains 92.2% of human performance on this task while scaling to analyze the entire traffic in 8 different languages of a large-scale conversational assistant. We present detailed ablation studies analyzing the impact of different modeling choices.

ReFinED: An Efficient Zero-shot-capable Approach to End-to-End Entity Linking
Tom Ayoola, Shubhi Tyagi, Joseph Fisher, Christos Christodoulopoulos and Andrea Pierleoni 08:45-09:00 (Quinault)
We introduce ReFinED, an efficient end-to-end entity linking model which uses fine-grained entity types and entity descriptions to perform linking. The model performs mention detection, fine-grained entity typing, and entity disambiguation for all mentions within a document in a single forward pass, making it more than 60 times faster than competitive existing approaches. ReFinED also surpasses state-of-the-art performance on standard entity linking datasets by an average of 3.7 F1. The model is capable of generalising to large-scale knowledge bases such as Wikidata (which has 15 times more entities than Wikipedia) and of zero-shot entity linking. The combination of speed, accuracy and scale makes ReFinED an effective and cost-efficient system for extracting entities from web-scale datasets, for which the model has been

successfully deployed.

In-person Poster Session 3

10:45-12:15 (Regency A & B)

SkillSpan: Hard and Soft Skill Extraction from English Job Postings
Mike Zhang, Kristian Nørgaard Jensen, Sif Dam Sonniks and Barbara Plank 10:45-12:15 (Regency A & B)
Skill Extraction (SE) is an important and widely-studied task useful to gain insights into labor market dynamics. However, there is a lacuna of datasets and annotation guidelines; available datasets are few and contain crowd-sourced labels on the span-level or labels from a predefined skill inventory. To address this gap, we introduce SKILLSPAN, a novel SE dataset consisting of 14.5K sentences and over 12.5K annotated spans. We release its respective guidelines created over three different sources annotated for hard and soft skills by domain experts. We introduce a BERT baseline (Devlin et al., 2019). To improve upon this baseline, we experiment with language models that are optimized for long spans (Joshi et al., 2020; Beltagy et al., 2020), continuous pre-training on the job posting domain (Han and Eisenstein, 2019; Gururangan et al., 2020), and multi-task learning (Caruana, 1997). Our results show that the domain-adapted models significantly outperform their non-adapted counterparts, and single-task outperforms multi-task learning.

Transparent Human Evaluation for Image Captioning
Jungo Kasai, Keisuke Sakaguchi, Lavinia Dunagan, Jacob Daniel Morrison, Ronan Le Bras, Yejin Choi and Noah Smith 10:45-12:15 (Regency A & B)
We establish THumB, a rubric-based human evaluation protocol for image captioning models. Our scoring rubrics and their definitions are carefully developed based on machine- and human-generated captions on the MSCOCO dataset. Each caption is evaluated along two main dimensions in a tradeoff (precision and recall) as well as other aspects that measure the text quality (fluency, conciseness, and inclusive language). Our evaluations demonstrate several critical problems of the current evaluation practice. Human-generated captions show substantially higher quality than machine-generated ones, especially in coverage of salient information (i.e., recall), while most automatic metrics say the opposite. Our rubric-based results reveal that CLIPScore, a recent metric that uses image features, better correlates with human judgments than conventional text-only metrics because it is more sensitive to recall. We hope that this work will promote a more transparent evaluation protocol for image captioning and its automatic metrics.

TVShowGuess: Character Comprehension in Stories as Speaker Guessing
Yisi Sang, Xiangyang Mou, Mo Yu, Shunyu Yao, Jing Li and Jeffrey Stanton 10:45-12:15 (Regency A & B)
We propose a new task for assessing machines' skills of understanding fictional characters in narrative stories. The task, TVShowGuess, builds on the scripts of TV series and takes the form of guessing the anonymous main characters based on the backgrounds of the scenes and the dialogues. Our human study supports that this form of task covers comprehension of multiple types of character persona, including understanding characters' personalities, facts and memories of personal experience, which are well aligned with the psychological and literary theories about the theory of mind (ToM) of human beings on understanding fictional characters during reading. We further propose new model architectures to support the contextualized encoding of long scene texts. Experiments show that our proposed approaches significantly outperform baselines, yet still largely lag behind the (nearly perfect) human performance. Our work serves as a first step toward the goal of narrative character comprehension.

CS1QA: A Dataset for Assisting Code-based Question Answering in an Introductory Programming Course
Changyoon Lee, Yeon Seonwoo and Alice Oh 10:45-12:15 (Regency A & B)
We introduce CS1QA, a dataset for code-based question answering in the programming education domain. CS1QA consists of 9,237 question-answer pairs gathered from chat logs in an introductory programming class using Python, and 17,698 unannotated chat data with code. Each question is accompanied with the student's code, and the portion of the code relevant to answering the question. We carefully design the annotation process to construct CS1QA, and analyze the collected dataset in detail. The tasks for CS1QA are to predict the question type, the relevant code snippet given the question and the code and retrieving an answer from the annotated corpus. Results for the experiments on several baseline models are reported and thoroughly analyzed. The tasks for CS1QA challenge models to understand both the code and natural language. This unique dataset can be used as a benchmark for source code comprehension and question answering in the educational setting.

Semantic Diversity in Dialogue with Natural Language Inference
Katherine Stasaski and Marti Hearst 10:45-12:15 (Regency A & B)
Generating diverse, interesting responses to chitchat conversations is a problem for neural conversational agents. This paper makes two substantial contributions to improving diversity in dialogue generation. First, we propose a novel metric which uses Natural Language Inference (NLI) to measure the semantic diversity of a set of model responses for a conversation. We evaluate this metric using an established framework (Tevet and Berant, 2021) and find strong evidence indicating NLI Diversity is correlated with semantic diversity. Specifically, we show that the contradiction relation is more useful than the neutral relation for measuring this diversity and that incorporating the NLI model's confidence achieves state-of-the-art results. Second, we demonstrate how to iteratively improve the semantic diversity of a sampled set of responses via a new generation procedure called Diversity Threshold Generation, which results in an average 137% increase in NLI Diversity compared to standard generation procedures.

Extending Multi-Text Sentence Fusion Resources via Pyramid Annotations
Daniela Brook Weiss, Paul Roit, Ori Ernst and Ido Dagan 10:45-12:15 (Regency A & B)
NLP models that process multiple texts often struggle in recognizing corresponding and salient information that is often differently phrased, and consolidating the redundancies across texts. To facilitate research of such challenges, the sentence fusion task was proposed, yet previous datasets for this task were very limited in their size and scope. In this paper, we revisit and substantially extend previous dataset creation efforts. With careful modifications, relabeling, and employing complementing data sources, we were able to more than triple the size of a notable earlier dataset. Moreover, we show that our extended version uses more representative texts for multi-document tasks and provides a more diverse training set, which substantially improves model performance.

ChapterBreak: A Challenge Dataset for Long-Range Language Models
Simeng Sun, Katherine Thai and Mohit Iyyer 10:45-12:15 (Regency A & B)
While numerous architectures for long-range language models (LRLMs) have recently been proposed, a meaningful evaluation of their discourse-level language understanding capabilities has not yet followed. To this end, we introduce ChapterBreak, a challenge dataset that provides an LRLM with a long segment from a narrative that ends at a chapter boundary and asks it to distinguish the beginning of the

ground-truth next chapter from a set of negative segments sampled from the same narrative. A fine-grained human annotation reveals that our dataset contains many complex types of chapter transitions (e.g., parallel narratives, cliffhanger endings) that require processing global context to comprehend. Experiments on ChapterBreak show that existing LRLMs fail to effectively leverage long-range context, substantially underperforming a segment-level model trained directly for this task. We publicly release our ChapterBreak dataset to spur more principled future research into LRLMs.

The USMLE® Step 2 Clinical Skills Patient Note Corpus
Victoria Yaneva, Janet Mee, Le An Ha, Polina Harik, Michael Jodoin and Alex J Mechaber 10:45-12:15 (Regency A & B)
This paper presents a corpus of 43,985 clinical patient notes (PNs) written by 35,156 examinees during the high-stakes USMLE® Step 2 Clinical Skills examination. In this exam, examinees interact with standardized patients - people trained to portray simulated scenarios called clinical cases. For each encounter, an examinee writes a PN, which is then scored by physician raters using a rubric of clinical concepts, expressions of which should be present in the PN. The corpus features PNs from 10 clinical cases, as well as the clinical concepts from the case rubrics. A subset of 2,840 PNs were annotated by 10 physician experts such that all 143 concepts from the case rubrics (e.g., shortness of breath) were mapped to 34,660 PN phrases (e.g., dyspnea, difficulty breathing). The corpus is available via a data sharing agreement with NBME and can be requested at https://www.nbme.org/services/data-sharing.

[TACL] **Czech Grammar Error Correction with a Large and Diverse Corpus**
Jakub Náplava, Milan Straka, Jana Straková and Alexandr Rosen 10:45-12:15 (Regency A & B)
We introduce a large and diverse Czech corpus annotated for grammatical error correction (GEC) with the aim to contribute to the still scarce data resources in this domain for languages other than English. The Grammar Error Correction Corpus for Czech (GECCC) offers a variety of four domains, covering error distributions ranging from high error density essays written by non-native speakers, to website texts, where errors are expected to be much less common. We compare several Czech GEC systems, including several Transformer-based ones, setting a strong baseline to future research. Finally, we meta-evaluate common GEC metrics against human judgements on our data. We make the new Czech GEC corpus publicly available under the CC BY-SA 4.0 license at http://hdl.handle.net/11234/1-4639 .

A Computational Acquisition Model for Multimodal Word Categorization
Uri Berger, Gabriel Stanovsky, Omri Abend and Lea Frermann 10:45-12:15 (Regency A & B)
Recent advances in self-supervised modeling of text and images open new opportunities for computational models of child language acquisition, which is believed to rely heavily on cross-modal signals. However, prior studies has been limited by their reliance on vision models trained on large image datasets annotated with a pre-defined set of depicted object categories. This is (a) not faithful to the information children receive and (b) prohibits the evaluation of such models with respect to category learning tasks, due to the pre-imposed category structure. We address this gap, and present a cognitively-inspired, multimodal acquisition model, trained from image-caption pairs on naturalistic data using cross-modal self-supervision. We show that the model learns word categories and object recognition abilities, and presents trends reminiscent of ones reported in the developmental literature.

[TACL] **He Thinks He Knows Better than the Doctors: BERT for Event Factuality Fails on Pragmatics**
Nanjiang Jiang and Marie-Catherine Hedwige de Marneffe 10:45-12:15 (Regency A & B)
We investigate how well BERT performs on predicting factuality in several existing English datasets, encompassing various linguistic constructions. Although BERT obtains a strong performance on most datasets, it does so by exploiting common surface patterns that correlate with certain factuality labels, and it fails on instances where pragmatic reasoning is necessary. Contrary to what the high performance suggests, we are still far from having a robust system for factuality prediction.

ProQA: Structural Prompt-based Pre-training for Unified Question Answering
Wanjun Zhong, Yifan Gao, Ning Ding, Yujia Qin, Zhiyuan Liu, Ming Zhou, Jiahai Wang, Jian Yin and Nan Duan 10:45-12:15 (Regency A & B)
Question Answering (QA) is a longstanding challenge in natural language processing. Existing QA works mostly focus on specific question types, knowledge domains, or reasoning skills. The specialty in QA research hinders systems from modeling commonalities between tasks and generalization for wider applications. To address this issue, we present ProQA, a unified QA paradigm that solves various tasks through a single model. ProQA takes a unified structural prompt as the bridge and improves the QA-centric ability by structural prompt-based pre-training. Through a structurally designed prompt-based input schema, ProQA concurrently models the knowledge generalization for all QA tasks while keeping the knowledge customization for every specific QA task. Furthermore, ProQA is pre-trained with structural prompt-formatted large-scale synthesized corpus, which empowers the model with the commonly-required QA ability. Experimental results on 11 QA benchmarks demonstrate that ProQA consistently boosts performance on both full data fine-tuning, few-shot learning, and zero-shot testing scenarios. Furthermore, ProQA exhibits strong ability in both continual learning and transfer learning by taking the advantages of the structural prompt.

DREAM: Improving Situational QA by First Elaborating the Situation
Yuling Gu, Bhavana Dalvi and Peter Clark 10:45-12:15 (Regency A & B)
When people answer questions about a specific situation, e.g., “I cheated on my mid-term exam last week. Was that wrong?”, cognitive science suggests that they form a mental picture of that situation before answering. While we do not know how language models (LMs) answer such questions, we conjecture that they may answer more accurately if they are also provided with additional details about the question situation, elaborating the “scene”. To test this conjecture, we train a new model, DREAM, to answer questions that elaborate the scenes that situated questions are about, and then provide those elaborations as additional context to a question-answering (QA) model. We find that DREAM is able to create better scene elaborations (more accurate, useful, and consistent) than a representative state-of-the-art, zero-shot model (Macaw). We also find that using the scene elaborations as additional context improves the answer accuracy of a downstream QA system, including beyond that obtainable by simply further fine-tuning the QA system on DREAM’s training data. These results suggest that adding focused elaborations about a situation can improve a system’s reasoning about it, and may serve as an effective way of injecting new scenario-based knowledge into QA models. Finally, our approach is dataset-neutral; we observe improved QA performance across different models, with even bigger gains on models with fewer parameters.

A New Concept of Knowledge based Question Answering (KBQA) System for Multi-hop Reasoning
Yu Wang, v.srinivasan@samsung.com v.srinivasan@samsung.com and Hongxia Jin 10:45-12:15 (Regency A & B)
Knowledge based question answering (KBQA) is a complex task for natural language understanding. Many KBQA approaches have been proposed in recent years, and most of them are trained based on labeled reasoning path. This hinders the system’s performance as many correct reasoning paths are not labeled as ground truth, and thus they cannot be learned. In this paper, we introduce a new concept of KBQA system which can leverage multiple reasoning paths’ information and only requires labeled answer as supervision. We name it as **M**utliple **R**easoning **P**aths KB**QA** System (MRP-QA). We conduct experiments on several benchmark datasets containing both single-hop simple questions as well as muti-hop complex questions, including WebQuestionSP (WQSP), ComplexWebQuestion-1.1 (CWQ), and PathQuestion-Large (PQL), and demonstrate strong performance.

Yes, No or IDK: The Challenge of Unanswerable Yes/No Questions
Elior Sulem, Jamaal Hay and Dan Roth 10:45-12:15 (Regency A & B)
The Yes/No QA task (Clark et al., 2019) consists of "Yes" or "No" questions about a given context. However, in realistic scenarios, the information provided in the context is not always sufficient in order to answer the question. For example, given the context "She married a lawyer from New-York.", we don't know whether the answer to the question "Did she marry in New York?" is "Yes" or "No". In this paper, we extend the Yes/No QA task, adding questions with an IDK answer, and show its considerable difficulty compared to the original 2-label task. For this purpose, we (i) enrich the BoolQ dataset (Clark et al., 2019) to include unanswerable questions and (ii) create out-of-domain test sets for the Yes/No/IDK QA task. We study the contribution of training on other Natural Language Understanding tasks. We focus in particular on Extractive QA (Rajpurkar et al., 2018) and Recognizing Textual Entailments (RTE; Dagan et al., 2013), analyzing the differences between 2 and 3 labels using the new data.

Long Context Question Answering via Supervised Contrastive Learning
Avi Caciularu, Ido Dagan, Jacob Goldberger and Arman Cohan 10:45-12:15 (Regency A & B)
Long-context question answering (QA) tasks require reasoning over a long document or multiple documents. Addressing these tasks often benefits from identifying a set of evidence spans (e.g., sentences), which provide supporting evidence for answering the question. In this work, we propose a novel method for equipping long-context QA models with an additional sequence-level objective for better identification of the supporting evidence. We achieve this via an additional contrastive supervision signal in finetuning, where the model is encouraged to explicitly discriminate supporting evidence sentences from negative ones by maximizing question-evidence similarity. The proposed additional loss exhibits consistent improvements on three different strong long-context transformer models, across two challenging question answering benchmarks – HotpotQA and QAsper.

[TACL] **Narrative Question Answering with Cutting-Edge Open-Domain QA Techniques: A Comprehensive Study**
Xiangyang Mou, Chenghao Yang, Mo Yu, Bingsheng Yao, Xiaoxiao Guo, Saloni Potdar and Hui Su 10:45-12:15 (Regency A & B)
Recent advancements in open-domain question answering (ODQA), i.e., finding answers from large open-domain corpus like Wikipedia, have led to human-level performance on many datasets. However, progress in QA over book stories (Book QA) lags behind despite its similar task formulation to ODQA. This work provides a comprehensive and quantitative analysis about the difficulty of Book QA: (1) We benchmark the research on the NarrativeQA dataset with extensive experiments with cutting-edge ODQA techniques. This quantifies the challenges Book QA poses, as well as advances the published state-of-the-art with a $\sim$7% absolute improvement on Rouge-L. (2) We further analyze the detailed challenges in Book QA through human studies.[1] Our findings indicate that the event-centric questions dominate this task, which exemplifies the inability of existing QA models to handle event-oriented scenarios.

DocAMR: Multi-Sentence AMR Representation and Evaluation
Tahira Naseem, Austin Blodgett, Sadhana Kumaravel, Tim O'Gorman, Young-Suk Lee, Jeffrey Flanigan, Ramón Fernandez Astudillo, Radu Florian, Salim Roukos and Nathan Schneider 10:45-12:15 (Regency A & B)
Despite extensive research on parsing of English sentences into Abstract Meaning Representation (AMR) graphs, which are compared to gold graphs via the Smatch metric, full-document parsing into a unified graph representation lacks well-defined representation and evaluation. Taking advantage of a super-sentential level of coreference annotation from previous work, we introduce a simple algorithm for deriving a unified graph representation, avoiding the pitfalls of information loss from over-merging and lack of coherence from under merging. Next, we describe improvements to the Smatch metric to make it tractable for comparing document-level graphs and use it to re-evaluate the best published document-level AMR parser. We also present a pipeline approach combining the top-performing AMR parser and coreference resolution systems, providing a strong baseline for future research.

CoSe-Co: Text Conditioned Generative CommonSense Contextualizer
Rachit Bansal, Milan Aggarwal, Sumit Bhatia, Jivat Neet Kaur and Balaji Krishnamurthy 10:45-12:15 (Regency A & B)
Pre-trained Language Models (PTLMs) have been shown to perform well on natural language tasks. Many prior works have leveraged structured commonsense present in the form of entities linked through labeled relations in Knowledge Graphs (KGs) to assist PTLMs. Retrieval approaches use KG as a separate static module which limits coverage since KGs contain finite knowledge. Generative methods train PTLMs on KG triples to improve the scale at which knowledge can be obtained. However, training on symbolic KG entities limits their applicability in tasks involving natural language text where they ignore overall context. To mitigate this, we propose a CommonSense Contextualizer (CoSe-Co) conditioned on sentences as input to make it generically usable in tasks for generating knowledge relevant to the overall context of input text. To train CoSe-Co, we propose a novel dataset comprising of sentence and commonsense knowledge pairs. The knowledge inferred by CoSe-Co is diverse and contain novel entities not present in the underlying KG. We augment generated knowledge in Multi-Choice QA and Open-ended CommonSense Reasoning tasks leading to improvements over current best methods on CSQA, ARC, QASC and OBQA datasets. We also demonstrate its applicability in improving performance of a baseline model for paraphrase generation task.

Batch-Softmax Contrastive Loss for Pairwise Sentence Scoring Tasks
Anton Chernyavskiy, Dmitry Ilvovsky, Pavel Kalinin and Preslav Nakov 10:45-12:15 (Regency A & B)
The use of contrastive loss for representation learning has become prominent in computer vision, and it is now getting attention in Natural Language Processing (NLP). Here, we explore the idea of using a batch-softmax contrastive loss when fine-tuning large-scale pre-trained transformer models to learn better task-specific sentence embeddings for pairwise sentence scoring tasks. We introduce and study a number of variations in the calculation of the loss as well as in the overall training procedure; in particular, we find that a special data shuffling can be quite important. Our experimental results show sizable improvements on a number of datasets and pairwise sentence scoring tasks including classification, ranking, and regression. Finally, we offer detailed analysis and discussion, which should be useful for researchers aiming to explore the utility of contrastive loss in NLP.

Label Definitions Improve Semantic Role Labeling
Li Zhang, Ishan Jindal and Yunyao Li 10:45-12:15 (Regency A & B)
Argument classification is at the core of Semantic Role Labeling. Given a sentence and the predicate, a semantic role label is assigned to each argument of the predicate. While semantic roles come with meaningful definitions, existing work has treated them as symbolic. Learning symbolic labels usually requires ample training data, which is frequently unavailable due to the cost of annotation. We instead propose to retrieve and leverage the definitions of these labels from the annotation guidelines. For example, the verb predicate "work" has arguments defined as "worker", "job", "employer", etc. Our model achieves state-of-the-art performance on the CoNLL09 dataset injected with label definitions given the predicate senses. The performance improvement is even more pronounced in low-resource settings when training data is scarce.

[1] `https://github.com/gorov/BookQA.`

Few-Shot Semantic Parsing with Language Models Trained on Code
Richard Shin and Benjamin Van Durme 10:45-12:15 (Regency A & B)
Large language models can perform semantic parsing with little training data, when prompted with in-context examples. It has been shown that this can be improved by formulating the problem as paraphrasing into canonical utterances, which casts the underlying meaning representation into a controlled natural language-like representation. Intuitively, such models can more easily output canonical utterances as they are closer to the natural language used for pre-training. Recently, models also pre-trained on code, like OpenAI Codex, have risen in prominence. For semantic parsing tasks where we map natural language into code, such models may prove more adept at it. In this paper, we test this hypothesis and find that Codex performs better on such tasks than equivalent GPT-3 models. We evaluate on Overnight and SMCalFlow and find that unlike GPT-3, Codex performs similarly when targeting meaning representations directly, perhaps because meaning representations are structured similar to code in these datasets.

Partial-input baselines show that NLI models can ignore context, but they don't.
Neha Srikanth and Rachel Rudinger 10:45-12:15 (Regency A & B)
When strong partial-input baselines reveal artifacts in crowdsourced NLI datasets, the performance of full-input models trained on such datasets is often dismissed as reliance on spurious correlations. We investigate whether state-of-the-art NLI models are capable of overriding default inferences made by a partial-input baseline. We introduce an evaluation set of 600 examples consisting of perturbed premises to examine a RoBERTa model's sensitivity to edited contexts. Our results indicate that NLI models are still capable of learning to condition on context—a necessary component of inferential reasoning—despite being trained on artifact-ridden datasets.

Improving negation detection with negation-focused pre-training
Thinh Hung Truong, Timothy Baldwin, Trevor Cohn and Karin Verspoor 10:45-12:15 (Regency A & B)
Negation is a common linguistic feature that is crucial in many language understanding tasks, yet it remains a hard problem due to diversity in its expression in different types of text. Recent works show that state-of-the-art NLP models underperform on samples containing negation in various tasks, and that negation detection models do not transfer well across domains. We propose a new negation-focused pre-training strategy, involving targeted data augmentation and negation masking, to better incorporate negation information into language models. Extensive experiments on common benchmarks show that our proposed approach improves negation detection performance and generalizability over the strong baseline NegBERT (Khandelwal and Sawant, 2020).

Paragraph-based Transformer Pre-training for Multi-Sentence Inference
Luca Di Liello, Siddhant Garg, Luca Soldaini and Alessandro Moschitti 10:45-12:15 (Regency A & B)
Inference tasks such as answer sentence selection (AS2) or fact verification are typically solved by fine-tuning transformer-based models as individual sentence-pair classifiers. Recent studies show that these tasks benefit from modeling dependencies across multiple candidate sentences jointly. In this paper, we first show that popular pre-trained transformers perform poorly when used for fine-tuning on multi-candidate inference tasks. We then propose a new pre-training objective that models the paragraph-level semantics across multiple input sentences. Our evaluation on three AS2 and one fact verification datasets demonstrates the superiority of our pre-training technique over the traditional ones for transformers used as joint models for multi-candidate inference tasks, as well as when used as cross-encoders for sentence-pair formulations of these tasks.

SUBS: Subtree Substitution for Compositional Semantic Parsing
Jingfeng Yang, Le Zhang and Diyi Yang 10:45-12:15 (Regency A & B)
Although sequence-to-sequence models often achieve good performance in semantic parsing for i.i.d. data, their performance is still inferior in compositional generalization. Several data augmentation methods have been proposed to alleviate this problem. However, prior work only leveraged superficial grammar or rules for data augmentation, which resulted in limited improvement. We propose to use subtree substitution for compositional data augmentation, where we consider subtrees with similar semantic functions as exchangeable. Our experiments showed that such augmented data led to significantly better performance on Scan and GeoQuery, and reached new SOTA on compositional split of GeoQuery.

Multi-Domain Targeted Sentiment Analysis
Orith Toledo-Ronen, Matan Orbach, Yoav Katz and Noam Slonim 10:45-12:15 (Regency A & B)
Targeted Sentiment Analysis (TSA) is a central task for generating insights from consumer reviews. Such content is extremely diverse, with sites like Amazon or Yelp containing reviews on products and businesses from many different domains. A real-world TSA system should gracefully handle that diversity. This can be achieved by a multi-domain model – one that is robust to the domain of the analyzed texts, and performs well on various domains. To address this scenario, we present a multi-domain TSA system based on augmenting a given training set with diverse weak labels from assorted domains. These are obtained through self-training on the Yelp reviews corpus. Extensive experiments with our approach on three evaluation datasets across different domains demonstrate the effectiveness of our solution. We further analyze how restrictions imposed on the available labeled data affect the performance, and compare the proposed method to the costly alternative of manually gathering diverse TSA labeled data. Our results and analysis show that our approach is a promising step towards a practical domain-robust TSA system.

UserIdentifier: Implicit User Representations for Simple and Effective Personalized Sentiment Analysis
*Fatemehsadat Mireshghallah, Vaishnavi Shrivastava, Milad Shokouhi, Taylor Berg-Kirkpatrick, Robert Sim and Dimitrios Dimitriadis*10:45-12:15 (Regency A & B)
Global models are typically trained to be as generalizable as possible. Invariance to the specific user is considered desirable since models are shared across multitudes of users. However, these models are often unable to produce personalized responses for individual users, based on their data. Contrary to widely-used personalization techniques based on few-shot and meta-learning, we propose UserIdentifier, a novel scheme for training a single shared model for all users. Our approach produces personalized responses by prepending a fixed, user-specific non-trainable string (called "user identifier") to each user's input text. Unlike prior work, this method doesn't need any additional model parameters, any extra rounds of personal few-shot learning or any change made to the vocabulary. We empirically study different types of user identifiers (numeric, alphanumeric, and also randomly generated) and demonstrate that, surprisingly, randomly generated user identifiers outperform the prefix-tuning based state-of-the-art approach by up to 13, on a suite of sentiment analysis datasets.

Data Augmentation with Dual Training for Offensive Span Detection
Nasim Nouri 10:45-12:15 (Regency A & B)
Recognizing offensive text is an important requirement for every content management system, especially for social networks. While the majority of the prior work formulate this problem as text classification, i.e., if a text excerpt is offensive or not, in this work we propose a novel model for offensive span detection (OSD), whose goal is to identify the spans responsible for the offensive tone of the text. One of the challenges to train a model for this novel setting is the lack of enough training data. To address this limitation, in this work we propose a novel method in which the large-scale pre-trained language model GPT-2 is employed to generate synthetic training data for OSD. In particular, we propose to train the GPT-2 model in a dual-training setting using the REINFORCE algorithm to generate in-domain, natural and diverse

training samples. Extensive experiments on the benchmark dataset for OSD reveal the effectiveness of the proposed method.

Analyzing Modality Robustness in Multimodal Sentiment Analysis
Devamanyu Hazarika, Yingting Li, Bo Cheng, Shuai Zhao, Roger Zimmermann and Soujanya Poria 10:45-12:15 (Regency A & B)
Building robust multimodal models are crucial for achieving reliable deployment in the wild. Despite its importance, less attention has been paid to identifying and improving the robustness of Multimodal Sentiment Analysis (MSA) models. In this work, we hope to address that by (i) Proposing simple diagnostic checks for modality robustness in a trained multimodal model. Using these checks, we find MSA models to be highly sensitive to a single modality, which creates issues in their robustness; (ii) We analyze well-known robust training strategies to alleviate the issues. Critically, we observe that robustness can be achieved without compromising on the original performance. We hope our extensive study–performed across five models and two benchmark datasets–and proposed procedures would make robustness an integral component in MSA research. Our diagnostic checks and robust training solutions are simple to implement and available at https://github.com/declare-lab/MSA-Robustness

Quantifying Language Variation Acoustically with Few Resources
Martijn Bartelds and Martijn Wieling 10:45-12:15 (Regency A & B)
Deep acoustic models represent linguistic information based on massive amounts of data. Unfortunately, for regional languages and dialects such resources are mostly not available. However, deep acoustic models might have learned linguistic information that transfers to low-resource languages. In this study, we evaluate whether this is the case through the task of distinguishing low-resource (Dutch) regional varieties. By extracting embeddings from the hidden layers of various wav2vec 2.0 models (including new models which are pre-trained and/or fine-tuned on Dutch) and using dynamic time warping, we compute pairwise pronunciation differences averaged over 10 words for over 100 individual dialects from four (regional) languages. We then cluster the resulting difference matrix in four groups and compare these to a gold standard, and a partitioning on the basis of comparing phonetic transcriptions. Our results show that acoustic models outperform the (traditional) transcription-based approach without requiring phonetic transcriptions, with the best performance achieved by the multilingual XLSR-53 model fine-tuned on Dutch. On the basis of only six seconds of speech, the resulting clustering closely matches the gold standard.

Session 5 - 14:15-15:45

Ethics, Bias, Fairness 1

14:15-15:45 (Columbia A)

What Factors Should Paper-Reviewer Assignments Rely On? Community Perspectives on Issues and Ideals in Conference Peer-Review
Terne Sasha Thorn Jakobsen and Anna Rogers 14:15-14:30 (Columbia A)
Both scientific progress and individual researcher careers depend on the quality of peer review, which in turn depends on paper-reviewer matching. Surprisingly, this problem has been mostly approached as an automated recommendation problem rather than as a matter where different stakeholders (area chairs, reviewers, authors) have accumulated experience worth taking into account. We present the results of the first survey of the NLP community, identifying common issues and perspectives on what factors should be considered by paper-reviewer matching systems. This study contributes actionable recommendations for improving future NLP conferences, and desiderata for interpretable peer review assignments.

Theory-Grounded Measurement of U.S. Social Stereotypes in English Language Models
Yang Trista Cao, Anna Sotnikova, Hal Daumé III, Rachel Rudinger and Linda Zou 14:30-14:45 (Columbia A)
NLP models trained on text have been shown to reproduce human stereotypes, which can magnify harms to marginalized groups when systems are deployed at scale. We adapt the Agency-Belief-Communion (ABC) stereotype model of Koch et al. (2016) from social psychology as a framework for the systematic study and discovery of stereotypic group-trait associations in language models (LMs). We introduce the sensitivity test (SeT) for measuring stereotypical associations from language models. To evaluate SeT and other measures using the ABC model, we collect group-trait judgments from U.S.-based subjects to compare with English LM stereotypes. Finally, we extend this framework to measure LM stereotyping of intersectional identities.

Benchmarking Intersectional Biases in NLP
John P. Lalor, Yi Yang, Kendall Smith, Nicole Forsgren and Ahmed Abbasi 14:45-15:00 (Columbia A)
There has been a recent wave of work assessing the fairness of machine learning models in general, and more specifically, on natural language processing (NLP) models built using machine learning techniques. While much work has highlighted biases embedded in state-of-the-art language models, and more recent efforts have focused on how to debias, research assessing the fairness and performance of biased/debiased models on downstream prediction tasks has been limited. Moreover, most prior work has emphasized bias along a single dimension such as gender or race. In this work, we benchmark multiple NLP models with regards to their fairness and predictive performance across a variety of NLP tasks. In particular, we assess intersectional bias - fairness across multiple demographic dimensions. The results show that while current debiasing strategies fare well in terms of the fairness-accuracy trade-off (generally preserving predictive power in debiased models), they are unable to effectively alleviate bias in downstream tasks. Furthermore, this bias is often amplified across dimensions (i.e., intersections). We conclude by highlighting possible causes and making recommendations for future NLP debiasing research.

Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection
Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi and Noah Smith 15:00-15:15 (Columbia A)
The perceived toxicity of language can vary based on someone's identity and beliefs, but this variation is often ignored when collecting toxic language datasets, resulting in dataset and model biases. We seek to understand the *who*, *why*, and *what* behind biases in toxicity annotations. In two online studies with demographically and politically diverse participants, we investigate the effect of annotator identities (*who*) and beliefs (*why*), drawing from social psychology research about hate speech, free speech, racist beliefs, political leaning, and more. We disentangle *what* is annotated as toxic by considering posts with three characteristics: anti-Black language, African American English (AAE) dialect, and vulgarity. Our results show strong associations between annotator identity and beliefs and their ratings of toxicity. Notably, more conservative annotators and those who scored highly on our scale for racist beliefs were less likely to rate anti-Black language as toxic, but more likely to rate AAE as toxic. We additionally present a case study illustrating how a popular toxicity detection system's ratings inherently reflect only specific beliefs and perspectives. Our findings call for contextualizing toxicity labels in social variables, which raises immense implications for toxic language annotation and detection.

Features or Spurious Artifacts? Data-centric Baselines for Fair and Robust Hate Speech Detection
Alan Ramponi and Sara Tonelli 15:15-15:30 (Columbia A)
Avoiding to rely on dataset artifacts to predict hate speech is at the cornerstone of robust and fair hate speech detection. In this paper we critically analyze lexical biases in hate speech detection via a cross-platform study, disentangling various types of spurious and authentic artifacts and analyzing their impact on out-of-distribution fairness and robustness. We experiment with existing approaches and propose simple yet surprisingly effective data-centric baselines. Our results on English data across four platforms show that distinct spurious artifacts require different treatments to ultimately attain both robustness and fairness in hate speech detection. To encourage research in this direction, we release all baseline models and the code to compute artifacts, pointing it out as a complementary and necessary addition to the data statements practice.

Gender Bias in Masked Language Models for Multiple Languages
Masahiro Kaneko, Aizhan Imankulova, Danushka Bollegala and Naoaki Okazaki 15:30-15:45 (Columbia A)
Masked Language Models (MLMs) pre-trained by predicting masked tokens on large corpora have been used successfully in natural language processing tasks for a variety of languages. Unfortunately, it was reported that MLMs also learn discriminative biases regarding attributes such as gender and race. Because most studies have focused on MLMs in English, the bias of MLMs in other languages has rarely been investigated. Manual annotation of evaluation data for languages other than English has been challenging due to the cost and difficulty in recruiting annotators. Moreover, the existing bias evaluation methods require the stereotypical sentence pairs consisting of the same context with attribute words (e.g. He/She is a nurse). We propose Multilingual Bias Evaluation (MBE) score, to evaluate bias in various languages using only English attribute word lists and parallel corpora between the target language and English without requiring manually annotated data. We evaluated MLMs in eight languages using the MBE and confirmed that gender-related biases are encoded in MLMs for all those languages. We manually created datasets for gender bias in Japanese and Russian to evaluate the validity of the MBE. The results show that the bias scores reported by the MBE significantly correlates with that computed from the above manually created datasets and the existing English datasets for gender bias.

Sentiment Analysis & Stylistic Analysis

14:15-15:45 (Columbia C)

Putting the Con in Context: Identifying Deceptive Actors in the Game of Mafia
Samee Omotayo Ibraheem, Gaoyue Zhou and John DeNero 14:15-14:30 (Columbia C)
While neural networks demonstrate a remarkable ability to model linguistic content, capturing contextual information related to a speaker's conversational role is an open area of research. In this work, we analyze the effect of speaker role on language use through the game of Mafia, in which participants are assigned either an honest or a deceptive role. In addition to building a framework to collect a dataset of Mafia game records, we demonstrate that there are differences in the language produced by players with different roles. We confirm that classification models are able to rank deceptive players as more suspicious than honest ones based only on their use of language. Furthermore, we show that training models on two auxiliary tasks outperforms a standard BERT-based text classification approach. We also present methods for using our trained models to identify features that distinguish between player roles, which could be used to assist players during the Mafia game.

CoMPM: Context Modeling with Speaker's Pre-trained Memory Tracking for Emotion Recognition in Conversation
Joosung Lee and Wooin Lee 14:30-14:45 (Columbia C)
As the use of interactive machines grow, the task of Emotion Recognition in Conversation (ERC) became more important. If the machine-generated sentences reflect emotion, more human-like sympathetic conversations are possible. Since emotion recognition in conversation is inaccurate if the previous utterances are not taken into account, many studies reflect the dialogue context to improve the performances. Many recent approaches show performance improvement by combining knowledge into modules learned from external structured data. However, structured data is difficult to access in non-English languages, making it difficult to extend to other languages. Therefore, we extract the pre-trained memory using the pre-trained language model as an extractor of external knowledge. We introduce CoMPM, which combines the speaker's pre-trained memory with the context model, and find that the pre-trained memory significantly improves the performance of the context model. CoMPM achieves the first or second performance on all data and is state-of-the-art among systems that do not leverage structured data. In addition, our method shows that it can be extended to other languages because structured knowledge is not required, unlike previous methods. Our code is available on github [2].

COGMEN: COntextualized GNN based Multimodal Emotion recognitioN
Abhinav Joshi, Ashwani Bhat, Ayush Jain, Atin Vikram Singh and Ashutosh Modi 14:45-15:00 (Columbia C)
Emotions are an inherent part of human interactions, and consequently, it is imperative to develop AI systems that understand and recognize human emotions. During a conversation involving various people, a person's emotions are influenced by the other speaker's utterances and their own emotional state over the utterances. In this paper, we propose COntextualized Graph Neural Network based Multi- modal Emotion recognitioN (COGMEN) system that leverages local information (i.e., inter/intra dependency between speakers) and global information (context). The proposed model uses Graph Neural Network (GNN) based architecture to model the complex dependencies (local and global information) in a conversation. Our model gives state-of-the- art (SOTA) results on IEMOCAP and MOSEI datasets, and detailed ablation experiments show the importance of modeling information at both levels.

Domain Confused Contrastive Learning for Unsupervised Domain Adaptation
Quanyu Long, Tianze Luo, Wenya Wang and Sinno Pan 15:00-15:15 (Columbia C)
In this work, we study Unsupervised Domain Adaptation (UDA) in a challenging self-supervised approach. One of the difficulties is how to learn task discrimination in the absence of target labels. Unlike previous literature which directly aligns cross-domain distributions or leverages reverse gradient, we propose Domain Confused Contrastive Learning (DCCL), which can bridge the source and target domains via domain puzzles, and retain discriminative representations after adaptation. Technically, DCCL searches for a most domain-challenging direction and exquisitely crafts domain confused augmentations as positive pairs, then it contrastively encourages the model to pull representations towards the other domain, thus learning more stable and effective domain invariances. We also investigate whether contrastive learning necessarily helps with UDA when performing other data augmentations. Extensive experiments demonstrate that DCCL significantly outperforms baselines, further ablation study and analysis also show the effectiveness and availability of DCCL.

Text Style Transfer via Optimal Transport
Nasim Nouri 15:15-15:30 (Columbia C)

[2] https://github.com/rungjoo/CoMPM

Text style transfer (TST) is a well-known task whose goal is to convert the style of the text (e.g., from formal to informal) while preserving its content. Recently, it has been shown that both syntactic and semantic similarities between the source and the converted text are important for TST. However, the interaction between these two concepts has not been modeled. In this work, we propose a novel method based on Optimal Transport for TST to simultaneously incorporate syntactic and semantic information into similarity computation between the source and the converted text. We evaluate the proposed method in both supervised and unsupervised settings. Our analysis reveal the superiority of the proposed model in both settings.

SSEGCN: Syntactic and Semantic Enhanced Graph Convolutional Network for Aspect-based Sentiment Analysis
Zheng Zhang, Zili Zhou and Yanna Wang 15:30-15:45 (Columbia C)
Aspect-based Sentiment Analysis (ABSA) aims to predict the sentiment polarity towards a particular aspect in a sentence. Recently, graph neural networks based on dependency tree convey rich structural information which is proven to be utility for ABSA. However, how to effectively harness the semantic and syntactic structure information from the dependency tree remains a challenging research question. In this paper, we propose a novel Syntactic and Semantic Enhanced Graph Convolutional Network (SSEGCN) model for ABSA task. Specifically, we propose an aspect-aware attention mechanism combined with self-attention to obtain attention score matrices of a sentence, which can not only learn the aspect-related semantic correlations, but also learn the global semantics of the sentence. In order to obtain comprehensive syntactic structure information, we construct syntactic mask matrices of the sentence according to the different syntactic distances between words. Furthermore, to combine syntactic structure and semantic information, we equip the attention score matrices by syntactic mask matrices. Finally, we enhance the node representations with graph convolutional network over attention score matrices for ABSA. Experimental results on benchmark datasets illustrate that our proposed model outperforms state-of-the-art methods.

Information Extraction 1

14:15-15:45 (Columbia D)

DEGREE: A Data-Efficient Generation-Based Event Extraction Model
I-Hung Hsu, Kuan-Hao Huang, Elizabeth Boschee, Scott Miller, Prem Natarajan, Kai-Wei Chang and Nanyun Peng 14:15-14:30 (Columbia D)
Event extraction requires high-quality expert human annotations, which are usually expensive. Therefore, learning a data-efficient event extraction model that can be trained with only a few labeled examples has become a crucial challenge. In this paper, we focus on low-resource end-to-end event extraction and propose DEGREE, a data-efficient model that formulates event extraction as a conditional generation problem. Given a passage and a manually designed prompt, DEGREE learns to summarize the events mentioned in the passage into a natural sentence that follows a predefined pattern. The final event predictions are then extracted from the generated sentence with a deterministic algorithm. DEGREE has three advantages to learn well with less training data. First, our designed prompts provide semantic guidance for DEGREE to leverage DEGREE and thus better capture the event arguments. Moreover, DEGREE is capable of using additional weakly-supervised information, such as the description of events encoded in the prompts. Finally, DEGREE learns triggers and arguments jointly in an end-to-end manner, which encourages the model to better utilize the shared knowledge and dependencies among them. Our experimental results demonstrate the strong performance of DEGREE for low-resource event extraction.

[TACL] **Text-based NP Enrichment**
Yanai Elazar, Victoria Basmov and Yoav Goldberg 14:30-14:45 (Columbia D)
Understanding the relations between entities denoted by NPs in a text is a critical part of human-like natural language understanding. However, only a fraction of such relations is covered by NLP tasks and models nowadays. In this work, we establish the task of text-based NP enrichment (TNE), that is, enriching each NP with all the preposition-mediated relations—both explicit and implicit—that hold between this and the other NPs in the text. The relations are represented as triplets, each denoted by two NPs related via a preposition. Humans recover such relations seamlessly, while current state-of-the-art models struggle with them due to the implicit nature of the problem. We build the first large-scale dataset for the problem, provide the formal framing and scope of annotation, analyze the data, and report the results of fine-tuned language models on the task, demonstrating the challenge it poses to current technology.

Few-Shot Document-Level Relation Extraction
Nicholas Popovic and Michael Färber 14:45-15:00 (Columbia D)
We present FREDo, a few-shot document-level relation extraction (FSDLRE) benchmark. As opposed to existing benchmarks which are built on sentence-level relation extraction corpora, we argue that document-level corpora provide more realism, particularly regarding none-of-the-above (NOTA) distributions. Therefore, we propose a set of FSDLRE tasks and construct a benchmark based on two existing supervised learning data sets, DocRED and sciERC. We adapt the state-of-the-art sentence-level method MNAV to the document-level and develop it further for improved domain adaptation. We find FSDLRE to be a challenging setting with interesting new characteristics such as the ability to sample NOTA instances from the support set. The data, code, and trained models are available online (https://github.com/nicpopovic/FREDo).

Joint Extraction of Entities, Relations, and Events via Modeling Inter-Instance and Inter-Label Dependencies
Minh Van Nguyen, Bonan Min, Franck Dernoncourt and Thien Huu Nguyen 15:00-15:15 (Columbia D)
Event trigger detection, entity mention recognition, event argument extraction, and relation extraction are the four important tasks in information extraction that have been performed jointly (Joint Information Extraction - JointIE) to avoid error propagation and leverage dependencies between the task instances (i.e., event triggers, entity mentions, relations, and event arguments). However, previous JointIE models often assume heuristic manually-designed dependency between the task instances and mean-field factorization for the joint distribution of instance labels, thus unable to capture optimal dependencies among instances and labels to improve representation learning and IE performance. To overcome these limitations, we propose to induce a dependency graph among task instances from data to boost representation learning. To better capture dependencies between instance labels, we propose to directly estimate their joint distribution via Conditional Random Fields. Noise Contrastive Estimation is introduced to address the maximization of the intractable joint likelihood for model training. Finally, to improve the decoding with greedy or beam search in prior work, we present Simulated Annealing to better find the globally optimal assignment for instance labels at decoding time. Experimental results show that our proposed model outperforms previous models on multiple IE tasks across 5 datasets and 2 languages.

Hyperbolic Relevance Matching for Neural Keyphrase Extraction
Mingyang Song, Yi Feng and Liping Jing 15:15-15:30 (Columbia D)
Keyphrase extraction is a fundamental task in natural language processing that aims to extract a set of phrases with important information from a source document. Identifying important keyphrases is the central component of keyphrase extraction, and its main challenge is learning to represent information comprehensively and discriminate importance accurately. In this paper, to address the above issues, we design

a new hyperbolic matching model (HyperMatch) to explore keyphrase extraction in hyperbolic space. Concretely, to represent information comprehensively, HyperMatch first takes advantage of the hidden representations in the middle layers of RoBERTa and integrates them as the word embeddings via an adaptive mixing layer to capture the hierarchical syntactic and semantic structures. Then, considering the latent structure information hidden in natural languages, HyperMatch embeds candidate phrases and documents in the same hyperbolic space via a hyperbolic phrase encoder and a hyperbolic document encoder. To discriminate importance accurately, HyperMatch estimates the importance of each candidate phrase by explicitly modeling the phrase-document relevance via the Poincaré distance and optimizes the whole model by minimizing the hyperbolic margin-based triplet loss. Extensive experiments are conducted on six benchmark datasets and demonstrate that HyperMatch outperforms the recent state-of-the-art baselines.

A Two-Stream AMR-enhanced Model for Document-level Event Argument Extraction
Runxin Xu, Peiyi Wang, Tianyu Liu, Shuang Zeng, Baobao Chang and Zhifang Sui 15:30-15:45 (Columbia D)
Most previous studies aim at extracting events from a single sentence, while document-level event extraction still remains under-explored. In this paper, we focus on extracting event arguments from an entire document, which mainly faces two critical problems: a) the long-distance dependency between trigger and arguments over sentences; b) the distracting context towards an event in the document. To address these issues, we propose a **T**wo-**S**tream **A**bstract meaning **R**epresentation enhanced extraction model (TSAR). TSAR encodes the document from different perspectives by a two-stream encoding module, to utilize local and global information and lower the impact of distracting context. Besides, TSAR introduces an AMR-guided interaction module to capture both intra-sentential and inter-sentential features, based on the locally and globally constructed AMR semantic graphs. An auxiliary boundary loss is introduced to enhance the boundary information for text spans explicitly. Extensive experiments illustrate that TSAR outperforms previous state-of-the-art by a large margin, with 2.54 F1 and 5.13 F1 performance gain on the public RAMS and WikiEvents datasets respectively, showing the superiority in the cross-sentence arguments extraction. We release our code in https://github.com/ PKUnlp-icler/TSAR.

Human-Centered NLP 1

14:15-15:45 (Elwha A)

Mapping the Design Space of Human-AI Interaction in Text Summarization
Ruijia Cheng, Alison Smith-Renner, Ke Zhang, Joel R. Tetreault and Alejandro Jaimes 14:15-14:30 (Elwha A)
Automatic text summarization systems commonly involve humans for preparing data or evaluating model performance, yet, there lacks a systematic understanding of humans' roles, experience, and needs when interacting with or being assisted by AI. From a human-centered perspective, we map the design opportunities and considerations for human-AI interaction in text summarization and broader text generation tasks. We first conducted a systematic literature review of 70 papers, developing a taxonomy of five interactions in AI-assisted text generation and relevant design dimensions. We designed text summarization prototypes for each interaction. We then interviewed 16 users, aided by the prototypes, to understand their expectations, experience, and needs regarding efficiency, control, and trust with AI in text summarization and propose design considerations accordingly.

Deconstructing NLG Evaluation: Evaluation Practices, Assumptions, and Their Implications
Kaitlyn Zhou, Su Lin Blodgett, Adam Trischler, Hal Daumé III, Kaheer Suleman and Alexandra Olteanu 14:30-14:45 (Elwha A)
There are many ways to express similar things in text, which makes evaluating natural language generation (NLG) systems difficult. Compounding this difficulty is the need to assess varying quality criteria depending on the deployment setting. While the landscape of NLG evaluation has been well-mapped, practitioners' goals, assumptions, and constraints—which inform decisions about what, when, and how to evaluate—are often partially or implicitly stated, or not stated at all. Combining a formative semi-structured interview study of NLG practitioners (N=18) with a survey study of a broader sample of practitioners (N=61), we surface goals, community practices, assumptions, and constraints that shape NLG evaluations, examining their implications and how they embody ethical considerations.

User-Centric Gender Rewriting
Bashar Alhafni, Nizar Habash and Houda Bouamor 14:45-15:00 (Elwha A)
In this paper, we define the task of gender rewriting in contexts involving two users (I and/or You) – first and second grammatical persons with independent grammatical gender preferences. We focus on Arabic, a gender-marking morphologically rich language. We develop a multi-step system that combines the positive aspects of both rule-based and neural rewriting models. Our results successfully demonstrate the viability of this approach on a recently created corpus for Arabic gender rewriting, achieving 88.42 M2 F0.5 on a blind test set. Our proposed system improves over previous work on the first-person-only version of this task, by 3.05 absolute increase in M2 F0.5. We demonstrate a use case of our gender rewriting system by using it to post-edit the output of a commercial MT system to provide personalized outputs based on the users' grammatical gender preferences. We make our code, data, and pretrained models publicly available.

Explaining Why: How Instructions and User Interfaces Impact Annotator Rationales When Labeling Text Data
Jamar L. Sullivan Jr., Will Brackenbury, Andrew McNutt, Kevin Bryson, Kwam Byll, Yuxin Chen, Michael Littman, Chenhao Tan and Blase Ur 15:00-15:15 (Elwha A)
In the context of data labeling, NLP researchers are increasingly interested in having humans select rationales, a subset of input tokens relevant to the chosen label. We conducted a 332-participant online user study to understand how humans select rationales, especially how different instructions and user interface affordances impact the rationales chosen. Participants labeled ten movie reviews as positive or negative, selecting words and phrases supporting their label as rationales. We varied the instructions given, the rationale-selection task, and the user interface. Participants often selected about 12% of input tokens as rationales, but selected fewer if unable to drag over multiple tokens at once. Whereas participants were near unanimous in their data labels, they were far less consistent in their rationales. The user interface affordances and task greatly impacted the types of rationales chosen. We also observed large variance across participants.

An Exploration of Post-Editing Effectiveness in Text Summarization
Vivian Lai, Alison Smith-Renner, Ke Zhang, Ruijia Cheng, Wenjuan Zhang, Joel R. Tetreault and Alejandro Jaimes 15:15-15:30 (Elwha A)
Automatic summarization methods are efficient but can suffer from low quality. In comparison, manual summarization is expensive but produces higher quality. Can humans and AI collaborate to improve summarization performance? In similar text generation tasks (e.g., machine translation), human-AI collaboration in the form of "post-editing" AI-generated text reduces human workload and improves the quality of AI output. Therefore, we explored whether post-editing offers advantages in text summarization. Specifically, we conducted an experiment with 72 participants, comparing post-editing provided summaries with manual summarization for summary quality, human efficiency, and user experience on formal (XSum news) and informal (Reddit posts) text. This study sheds valuable insights on when post-editing is useful for text summarization: it helped in some cases (e.g., when participants lacked domain knowledge) but not in others (e.g., when provided summaries include inaccurate information). Participants' different editing strategies and needs for assistance offer implications for future

human-AI summarization systems.

The Why and The How: A Survey on Natural Language Interaction in Visualization
Henrik Voigt, Ozge Alacam, Monique Meuschke, Kai Lawonn and Sina Zarrieß 15:30-15:45 (Elwha A)
Natural language as a modality of interaction is becoming increasingly popular in the field of visualization. In addition to the popular query interfaces, other language-based interactions such as annotations, recommendations, explanations, or documentation experience growing interest. In this survey, we provide an overview of natural language-based interaction in the research area of visualization. We discuss a renowned taxonomy of visualization tasks and classify 119 related works to illustrate the state-of-the-art of how current natural language interfaces support their performance. We examine applied NLP methods and discuss human-machine dialogue structures with a focus on initiative, duration, and communicative functions in recent visualization-oriented dialogue interfaces. Based on this overview, we point out interesting areas for the future application of NLP methods in the field of visualization.

NLP Applications 3

14:15-15:45 (Elwha B)

Cryptocurrency Bubble Detection: A New Stock Market Dataset, Financial Task & Hyperbolic Models
Ramit Sawhney, Shivam Agarwal, Vivek Mittal, Paolo Rosso, Vikram Nanda and Sudheer Chava 14:15-14:30 (Elwha B)
The rapid spread of information over social media influences quantitative trading and investments. The growing popularity of speculative trading of highly volatile assets such as cryptocurrencies and meme stocks presents a fresh challenge in the financial realm. Investigating such "bubbles" - periods of sudden anomalous behavior of markets are critical in better understanding investor behavior and market dynamics. However, high volatility coupled with massive volumes of chaotic social media texts, especially for underexplored assets like cryptocoins pose a challenge to existing methods. Taking the first step towards NLP for cryptocoins, we present and publicly release CryptoBubbles, a novel multi- span identification task for bubble detection, and a dataset of more than 400 cryptocoins from 9 exchanges over five years spanning over two million tweets. Further, we develop a set of sequence-to-sequence hyperbolic models suited to this multi-span identification task based on the power-law dynamics of cryptocurrencies and user behavior on social media. We further test the effectiveness of our models under zero-shot settings on a test set of Reddit posts pertaining to 29 “meme stocks”, which see an increase in trade volume due to social media hype. Through quantitative, qualitative, and zero-shot analyses on Reddit and Twitter spanning cryptocoins and meme-stocks, we show the practical applicability of CryptoBubbles and hyperbolic models.

Many Hands Make Light Work: Using Essay Traits to Automatically Score Essays
Rahul Kumar, Sandeep Mathias, Sriparna Saha and Pushpak Bhattacharyya 14:30-14:45 (Elwha B)
Most research in the area of automatic essay grading (AEG) is geared towards scoring the essay *holistically* while there has also been little work done on scoring individual essay traits. In this paper, we describe a way to score essays using a multi-task learning (MTL) approach, where scoring the essay holistically is the primary task, and scoring the essay traits is the auxiliary task. We compare our results with a single-task learning (STL) approach, using both LSTMs and BiLSTMs. To find out which traits work best for different types of essays, we conduct ablation tests for each of the essay traits. We also report the runtime and number of training parameters for each system. We find that MTL-based BiLSTM system gives the best results for scoring the essay holistically, as well as performing well on scoring the essay traits. The MTL systems also give a speed-up of between **2.30** to **3.70** times the speed of the STL system, when it comes to scoring the essay and all the traits.

Aligning to Social Norms and Values in Interactive Narratives
Prithviraj Ammanabrolu, Liwei Jiang, Maarten Sap, Hannaneh Hajishirzi and Yejin Choi 14:45-15:00 (Elwha B)
We focus on creating agents that act in alignment with socially beneficial norms and values in interactive narratives or text-based games—environments wherein an agent perceives and interacts with a world through natural language. Such interactive agents are often trained via reinforcement learning to optimize task performance, even when such rewards may lead to agent behaviors that violate societal norms—causing harm either to the agent itself or other entities in the environment. Social value alignment refers to creating agents whose behaviors conform to expected moral and social norms for a given context and group of people—in our case, it means agents that behave in a manner that is less harmful and more beneficial for themselves and others.

We build on the Jiminy Cricket benchmark (Hendrycks et al. 2021), a set of 25 annotated interactive narratives containing thousands of morally salient scenarios covering everything from theft and bodily harm to altruism. We introduce the GALAD (Game-value ALignment through Action Distillation) agent that uses the social commonsense knowledge present in specially trained language models to contextually restrict its action space to only those actions that are aligned with socially beneficial values. An experimental study shows that the GALAD agent makes decisions efficiently enough to improve state-of-the-art task performance by 4% while reducing the frequency of socially harmful behaviors by 25% compared to strong contemporary value alignment approaches.

LITE: Intent-based Task Representation Learning Using Weak Supervision
Naoki Otani, Michael Gamon, Sujay Kumar Jauhar, Mei Yang, Sri Raghu Malireddi and Oriana Riva 15:00-15:15 (Elwha B)
Users write to-dos as personal notes to themselves, about things they need to complete, remember or organize. To-do texts are usually short and under-specified, which poses a challenge for current text representation models. Yet, understanding and representing their meaning is the first step towards providing intelligent assistance for to-do management. We address this problem by proposing a neural multi-task learning framework, LITE, which extracts representations of English to-do tasks with a multi-head attention mechanism on top of a pre-trained text encoder. To adapt representation models to to-do texts, we collect weak-supervision labels from semantically rich external resources (e.g., dynamic commonsense knowledge bases), following the principle that to-do tasks with similar intents have similar labels. We then train the model on multiple generative/predictive training objectives jointly. We evaluate our representation model on four downstream tasks and show that our approach consistently improves performance over baseline models, achieving error reduction of up to 38.7%.

Forecasting COVID-19 Caseloads Using Unsupervised Embedding Clusters of Social Media Posts
Felix Drinkall, Stefan Zohren and Janet B. Pierrehumbert 15:15-15:30 (Elwha B)
We present a novel approach incorporating transformer-based language models into infectious disease modelling. Text-derived features are quantified by tracking high-density clusters of sentence-level representations of Reddit posts within specific US states' COVID-19 subreddits. We benchmark these clustered embedding features against features extracted from other high-quality datasets. In a threshold-classification task, we show that they outperform all other feature types at predicting upward trend signals, a significant result for infectious disease modelling in areas where epidemiological data is unreliable. Subsequently, in a time-series forecasting task, we fully utilise the predictive power of the caseload and compare the relative strengths of using different supplementary datasets as covariate feature sets in a transformer-based time-series model.

Context-Aware Abbreviation Expansion Using Large Language Models
*Shanqing Cai, Subhashini Venugopalan, Katrin Tomanek, Ajit Narayanan, Meredith Ringel Morris and Michael Brenner*15:30-15:45 (Elwha B)
Motivated by the need for accelerating text entry in augmentative and alternative communication (AAC) for people with severe motor impairments, we propose a paradigm in which phrases are abbreviated aggressively as primarily word-initial letters. Our approach is to expand the abbreviations into full-phrase options by leveraging conversation context with the power of pretrained large language models (LLMs). Through zero-shot, few-shot, and fine-tuning experiments on four public conversation datasets, we show that for replies to the initial turn of a dialog, an LLM with 64B parameters is able to exactly expand over 70% of phrases with abbreviation length up to 10, leading to an effective keystroke saving rate of up to about 77% on these exact expansions. Including a small amount of context in the form of a single conversation turn more than doubles abbreviation expansion accuracies compared to having no context, an effect that is more pronounced for longer phrases. Additionally, the robustness of models against typo noise can be enhanced through fine-tuning on noisy data.

SRW In-Person Poster Session

14:15-15:45 (Regency A & B)

Systematicity Emerges in Transformers when Abstract Grammatical Roles Guide Attention
Ayush K Chakravarthy, Jacob Labe Russin and Randall O'Reilly 15:45-14:15 (Regency A & B)
Systematicity is thought to be a key inductive bias possessed by humans that is lacking in standard natural language processing systems such as those utilizing transformers. In this work, we investigate the extent to which the failure of transformers on systematic generalization tests can be attributed to a lack of linguistic abstraction in its attention mechanism. We develop a novel modification to the transformer by implementing two separate input streams: a role stream controls the attention distributions (i.e., queries and keys) at each layer, and a filler stream determines the values. Our results show that when abstract role labels are assigned to input sequences and provided to the role stream, systematic generalization is improved.

Grounding in social media: An approach to building a chit-chat dialogue model
Ritvik Choudhary and Daisuke Kawahara 15:45-14:15 (Regency A & B)
Building open-domain dialogue systems capable of rich human-like conversational ability is one of the fundamental challenges in language generation. However, even with recent advancements in the field, existing open-domain generative models fail to capture and utilize external knowledge, leading to repetitive or generic responses to unseen utterances. Current work on knowledge-grounded dialogue generation primarily focuses on persona incorporation or searching a fact-based structured knowledge source such as Wikipedia. Our method takes a broader and simpler approach, which aims to improve the raw conversation ability of the system by mimicking the human response behavior through casual interactions found on social media. Utilizing a joint retriever-generator setup, the model queries a large set of filtered comment data from Reddit to act as additional context for the seq2seq generator. Automatic and human evaluations on open-domain dialogue datasets demonstrate the effectiveness of our approach.

ExtraPhrase: Efficient Data Augmentation for Abstractive Summarization
Mengsay Loem, Sho Takase, Masahiro Kaneko and Naoaki Okazaki 15:45-14:15 (Regency A & B)
Neural models trained with large amount of parallel data have achieved impressive performance in abstractive summarization tasks. However, large-scale parallel corpora are expensive and challenging to construct. In this work, we introduce a low-cost and effective strategy, ExtraPhrase, to augment training data for abstractive summarization tasks. ExtraPhrase constructs pseudo training data in two steps: extractive summarization and paraphrasing. We extract major parts of an input text in the extractive summarization step and obtain its diverse expressions with the paraphrasing step. Through experiments, we show that ExtraPhrase improves the performance of abstractive summarization tasks by more than 0.50 points in ROUGE scores compared to the setting without data augmentation. ExtraPhrase also outperforms existing methods such as back-translation and self-training. We also show that ExtraPhrase is significantly effective when the amount of genuine training data is remarkably small, i.e., a low-resource setting. Moreover, ExtraPhrase is more cost-efficient than the existing approaches

Neural Retriever and Go Beyond: A Thesis Proposal
Man Luo 15:45-14:15 (Regency A & B)
Information Retriever (IR) aims to find the relevant documents (e.g. snippets, passages, and articles) to a given query at large scale. IR plays an important role in many tasks such as open domain question answering and dialogue systems, where external knowledge is needed. In the past, searching algorithms based on term matching have been widely used. Recently, neural-based algorithms (termed as neural retrievers) have gained more attention which can mitigate the limitations of traditional methods. Regardless of the success achieved by neural retrievers, they still face many challenges, e.g. suffering from a small amount of training data and failing to answer simple entity-centric questions. Furthermore, most of the existing neural retrievers are developed for pure-text query. This prevents them from handling multi-modality queries (i.e. the query is composed of textual description and images). This proposal has two goals. First, we introduce methods to address the abovementioned issues of neural retrievers from three angles, new model architectures, IR-oriented pretraining tasks, and generating large scale training data. Second, we identify the future research direction and propose potential corresponding solution.

Improving Classification of Infrequent Cognitive Distortions: Domain-Specific Model vs. Data Augmentation
Xiruo Ding, Kevin Lybarger, Justin Tauscher and Trevor Cohen 15:45-14:15 (Regency A & B)
Cognitive distortions are counterproductive patterns of thinking that are one of the targets of cognitive behavioral therapy (CBT). These can be challenging for clinicians to detect, especially those without extensive CBT training or supervision. Text classification methods can approximate expert clinician judgment in the detection of frequently occurring cognitive distortions in text-based therapy messages. However, performance with infrequent distortions is relatively poor. In this study, we address this sparsity problem with two approaches: Data Augmentation and Domain-Specific Model. The first approach includes Easy Data Augmentation, back translation, and mixup techniques. The second approach utilizes a domain-specific pretrained language model, MentalBERT. To examine the viability of different data augmentation methods, we utilized a real-world dataset of texts between therapists and clients diagnosed with serious mental illness that was annotated for distorted thinking. We found that with optimized parameter settings, mixup was helpful for rare classes. Performance improvements with an augmented model, MentalBERT, exceed those obtained with data augmentation.

What "Drives" the Use of Metaphorical Language? Negative Insights from Abstractness, Affect, Discourse Coherence and Contextualized Word Representations
Prisca Piccirilli and Sabine Schulte im Walde 15:45-14:15 (Regency A & B)
Which features in a specific discourse trigger the use of metaphorical language, rather than literal alternatives? Many NLP approaches to

metaphor rely on cognitive and (psycho)linguistic insights and have successfully applied models of discourse coherence, abstractness and affect, but this study indicates that these properties do not systematically explain metaphorical vs. literal preferences.

Generate, Evaluate, and Select: A Dialogue System with a Response Evaluator for Diversity-Aware Response Generation
Ryoma Sakaeda and Daisuke Kawahara 15:45-14:15 (Regency A & B)
We aim to overcome the lack of diversity in responses of current dialogue systems and to develop a dialogue system that is engaging as a conversational partner. We propose a generator-evaluator model that evaluates multiple responses generated by a response generator and selects the best response by an evaluator. By generating multiple responses, we obtain diverse responses. We conduct human evaluations to compare the output of the proposed system with that of a baseline system. The results of the human evaluations showed that the proposed system's responses were often judged to be better than the baseline system's, and indicated the effectiveness of the proposed method.

Building a Personalized Dialogue System with Prompt-Tuning
Tomohito Kasahara, Daisuke Kawahara, Nguyen Tung, Shengzhe Li, Kenta Shinzato and Toshinori Sato 15:45-14:15 (Regency A & B)
Dialogue systems without consistent responses are not attractive. In this study, we build a dialogue system that can respond based on a given character setting (persona) to bring consistency. Considering the trend of the rapidly increasing scale of language models, we propose an approach that uses prompt-tuning, which has low learning costs, on pre-trained large-scale language models. The results of the automatic and manual evaluations in English and Japanese show that it is possible to build a dialogue system with more natural and personalized responses with less computational resources than fine-tuning.

MM-GATBT: Enriching Multimodal Representation Using Graph Attention Network
Seung Byum Seo, Hyoungwook Nam and Payam Delgosha 15:45-14:15 (Regency A & B)
While there have been advances in Natural Language Processing (NLP), their success is mainly gained by applying a self-attention mechanism into single or multi-modalities. While this approach has brought significant improvements in multiple downstream tasks, it fails to capture the interaction between different entities. Therefore, we propose MM-GATBT, a multimodal graph representation learning model that captures not only the relational semantics within one modality but also the interactions between different modalities. Specifically, the proposed method constructs image-based node embedding which contains relational semantics of entities. Our empirical results show that MM-GATBT achieves state-of-the-art results among all published papers on the MM-IMDb dataset.

ViT5: Pretrained Text-to-Text Transformer for Vietnamese Language Generation
Long Phan, Hieu Tran, Hieu Nguyen and Trieu H. Trinh 15:45-14:15 (Regency A & B)
We present ViT5, a pretrained Transformer-based encoder-decoder model for the Vietnamese language. With T5-style self-supervised pretraining, ViT5 is trained on a large corpus of high-quality and diverse Vietnamese texts. We benchmark ViT5 on two downstream text generation tasks, Abstractive Text Summarization and Named Entity Recognition. Although Abstractive Text Summarization has been widely studied for the English language thanks to its rich and large source of data, there has been minimal research into the same task in Vietnamese, a much lower resource language. In this work, we perform exhaustive experiments on both Vietnamese Abstractive Summarization and Named Entity Recognition, validating the performance of ViT5 against many other pretrained Transformer-based encoder-decoder models. Our experiments show that ViT5 significantly outperforms existing models and achieves state-of-the-art results on Vietnamese Text Summarization. On the task of Named Entity Recognition, ViT5 is competitive against previous best results from pretrained encoder-based Transformer models. Further analysis shows the importance of context length during the self-supervised pretraining on downstream performance across different settings.

Compositional Generalization in Grounded Language Learning via Induced Model Sparsity
Sam Spilsbury and Alexander Ilin 15:45-14:15 (Regency A & B)
We provide a study of how induced model sparsity can help achieve compositional generalization and better sample efficiency in grounded language learning problems. We consider simple language-conditioned navigation problems in a grid world environment with disentangled observations. We show that standard neural architectures do not always yield compositional generalization. To address this, we design an agent that contains a goal identification module that encourages sparse correlations between words in the instruction and attributes of objects, composing them together to find the goal. The output of the goal identification module is the input to a value iteration network planner. Our agent maintains a high level of performance on goals containing novel combinations of properties even when learning from a handful of demonstrations. We examine the internal representations of our agent and find the correct correspondences between words in its dictionary and attributes in the environment.

How do people talk about images? A study on open-domain conversations with images.
Yi-Pei Chen, Nobuyuki Shimizu, Takashi Miyazaki and Hideki Nakayama 15:45-14:15 (Regency A & B)
This paper explores how humans conduct conversations with images by investigating an open-domain image conversation dataset, ImageChat. We examined the conversations with images from the perspectives of $extitimagerelevancy$ and $extitimageinformation.$ We found that utterances/conversations are not always related to the given image, and conversation topics diverge within three turns about half of the time. Besides image objects, more comprehensive non-object image information is also indispensable. After inspecting the causes, we suggested that understanding the overall scenario of image and connecting objects based on their high-level attributes might be very helpful to generate more engaging open-domain conversations when an image is presented. We proposed enriching the image information with image caption and object tags based on our analysis. With our proposed $extitimage^+$ features, we improved automatic metrics including BLEU and Bert Score, and increased the diversity and image-relevancy of generated responses to the strong baseline. The result verifies that our analysis provides valuable insights and could facilitate future research on open-domain conversations with images.

Preschool Children Speech Recognition for Early Childhood Intervention: Motivation and Challenges
Satwik Dutta, Dwight W. Irvin and John H. L. Hansen 15:45-14:15 (Regency A & B)
Monitoring child development in terms of speech/language skills has a long-term impact on their overall growth. As student diversity continue to expand in US classrooms, there is a growing need to benchmark social engagement, both from a teacher-student perspective, as well as student-student content. Given various challenges with direct observation, deploying speech technology can assist in extracting meaningful information for teachers. These will help teachers to identify and respond to students in need, immediately impacting their early learning and interest. This study takes a deep dive into exploring hybrid ASR solutions for low-resource spontaneous preschool (3-5yrs) children (with and without developmental delays) speech, being involved in various activities, and interacting with teachers and peers in naturalistic classrooms. For the purpose of data augmentation, various out-of-domain corpora over a wide and limited age range, both scripted and spontaneous were considered. Acoustic models based on factorized time-delay neural networks, and both N-gram and neural language models were considered. Results indicate that young children have significantly different/developing articulation skills as compared to older children. Out-of-domain transcripts of interactions between young children and adults however enhances language model performance. Overall transcription of such data, including various non-linguistic markers, poses additional challenges.

A Simple Approach to Jointly Rank Passages and Select Relevant Sentences in the OBQA Context

Man Luo, Shuguang Chen and Chitta Baral 15:45-14:15 (Regency A & B)
In the open book question answering (OBQA) task, selecting the relevant passages and sentences from distracting information is crucial to reason the answer to a question. HotpotQA dataset is designed to teach and evaluate systems to do both passage ranking and sentence selection. Many existing frameworks use separate models to select relevant passages and sentences respectively. Such systems not only have high complexity in terms of the parameters of models but also fail to take the advantage of training these two tasks together since one task can be beneficial for the other one. In this work, we present a simple yet effective framework to address these limitations by jointly ranking passages and selecting sentences. Furthermore, we propose consistency and similarity constraints to promote the correlation and interaction between passage ranking and sentence selection.The experiments demonstrate that our framework can achieve competitive results with previous systems and outperform the baseline by 28

Multimodal Modeling of Task-Mediated Confusion
Camille Mince, Skye Rhomberg, Cecilia Alm, Reynold Bailey and Alex Ororbia 15:45-14:15 (Regency A & B)
In order to build more human-like cognitive agents, systems capable of detecting various human emotions must be designed to respond appropriately. Confusion, the combination of an emotional and cognitive state, is under-explored. In this paper, we build upon prior work to develop models that detect confusion from three modalities: video (facial features), audio (prosodic features), and text (transcribed speech features). Our research improves the data collection process by allowing for continuous (as opposed to discrete) annotation of confusion levels. We also craft models based on recurrent neural networks (RNNs) given their ability to predict sequential data. In our experiments, we find that text and video modalities are the most important in predicting confusion while the explored audio features are relatively unimportant predictors of confusion in our data.

Machine Narrative Comprehension in Fictional Characters Personality Prediction Task
Yisi Sang, Xiangyang Mou, Mo Yu, Dakuo Wang, Jing Li and Jeffrey Stanton 15:45-14:15 (Regency A & B)
An NLP model that understands stories should also be able to understand the characters, which is underexplored till now. To support the development of neural models for this purpose, we construct a benchmark, Story2Personality. The task is to predict a movie character's personality based on the narratives. Experiments show that our task is challenging for the existing text classification models, as none is able to largely outperform random guesses. We then proposed a multi-view model for personality prediction using both verbal and non-verbal descriptions, which significantly improved the performance. The uniqueness and challenges in our dataset call for the development of narrative comprehension techniques from the perspective of understanding characters.

Divide & Conquer for Entailment-aware Multi-hop Evidence Retrieval
Fan Luo and Mihai Surdeanu 15:45-14:15 (Regency A & B)
Lexical and semantic matches are commonly used as relevance measurements for information retrieval. Together they estimate the semantic equivalence between the query and the candidates. However, semantic equivalence is not the only relevance signal need to be considered when retrieving evidences for multi-hop questions. In this work, we demonstrate that textual entailment relation is another important relevance dimension that should be considered. To retrieve evidences that are either semantic equivalent to or entailed by the question simultaneously, we divide the task of evidence retrieval for multi-hop question answering (QA) into two sub-tasks, i.e., semantic textual similarity and inference similarity retrieval. We propose two ensemble models, EAR and EARnest, which tackle each of the sub-tasks separately with off-the-shelf retrieval models, and jointly retrieve sentences with the consideration of the diverse relevance signals. Experimental results on HotpotQA verify that our models not only significantly outperform all the single retrieval models it based on, but also more effective than two intuitive ensemble baseline models.

Neural Networks in a Product of Hyperbolic Spaces
Jun Takeuchi, Noriki Nishida and Hideki Nakayama 15:45-14:15 (Regency A & B)
Machine learning in hyperbolic spaces has attracted much attention in natural language processing and many other fields. In particular, Hyperbolic Neural Networks (HNNs) have improved a wide variety of tasks, from machine translation to knowledge graph embedding. Although some studies have reported the effectiveness of embedding into the product of multiple hyperbolic spaces, HNNs have mainly been constructed in a single hyperbolic space, and their extension to product spaces has not been sufficiently studied. Therefore, we propose a novel method to extend a given HNN in a single space to a product of hyperbolic spaces. We apply our method to Hyperbolic Graph Convolutional Networks (HGCNs), extending several HNNs. Our model improved the graph node classification accuracy especially on datasets with tree-like structures. The results suggest that neural networks in a product of hyperbolic spaces can be more effective than in a single space in representing structural data.

Strong Heuristics for Named Entity Linking
Marko Čuljak, Andreas Spitz, Robert West and Akhil Arora 15:45-14:15 (Regency A & B)
Named entity linking (NEL) in news is a challenging endeavour due to the frequency of unseen and emerging entities, which necessitates the use of unsupervised or zero-shot methods. However, such methods tend to come with caveats, such as no integration of suitable knowledge bases (like Wikidata) for emerging entities, a lack of scalability, and poor interpretability. Here, we consider person disambiguation in Quotebank, a massive corpus of speaker-attributed quotations from the news, and investigate the suitability of intuitive, lightweight, and scalable heuristics for NEL in web-scale corpora. Our best performing heuristic disambiguates 94% and 63% of the mentions on Quotebank and the AIDA-CoNLL benchmark, respectively. Additionally, the proposed heuristics compare favourably to the state-of-the-art unsupervised and zero-shot methods, Eigenthemes and mGENRE, respectively, thereby serving as strong baselines for unsupervised and zero-shot entity linking.

Unifying Parsing and Tree-Structured Models for Generating Sentence Semantic Representations
Antoine Simoulin and Benoit Crabbé 15:45-14:15 (Regency A & B)
We introduce a novel tree-based model that learns its composition function together with its structure. The architecture produces sentence embeddings by composing words according to an induced syntactic tree. The parsing and the composition functions are explicitly connected and, therefore, learned jointly. As a result, the sentence embedding is computed according to an interpretable linguistic pattern and may be used on any downstream task. We evaluate our encoder on downstream tasks, and we observe that it outperforms tree-based models relying on external parsers. In some configurations, it is even competitive with Bert base model. Our model is capable of supporting multiple parser architectures. We exploit this property to conduct an ablation study by comparing different parser initializations. We explore to which extent the trees produced by our model compare with linguistic structures and how this initialization impacts downstream performances. We empirically observe that downstream supervision troubles producing stable parses and preserving linguistically relevant structures.

Defending Compositionality in Emergent Languages
Michal Auersperger and Pavel Pecina 15:45-14:15 (Regency A & B)
Compositionality has traditionally been understood as a major factor in productivity of language and, more broadly, human cognition. Yet, recently some research started to question its status showing that artificial neural networks are good at generalization even without noticeable compositional behavior. We argue some of these conclusions are too strong and/or incomplete. In the context of a two-agent communication game, we show that compositionality indeed seems essential for successful generalization when the evaluation is done on a suitable dataset.

Exploring the Effect of Dialect Mismatched Language Models in Telugu Automatic Speech Recognition
Aditya Yadavalli, Ganesh Sai Mirishkar and Anil Vuppala 15:45-14:15 (Regency A & B)
Previous research has found that Acoustic Models (AM) of an Automatic Speech Recognition (ASR) system are susceptible to dialect variations within a language, thereby adversely affecting the ASR. To counter this, researchers have proposed to build a dialect-specific AM while keeping the Language Model (LM) constant for all the dialects. This study explores the effect of dialect mismatched LM by considering three different Telugu regional dialects: Telangana, Coastal Andhra, and Rayalaseema. We show that dialect variations that surface in the form of a different lexicon, grammar, and occasionally semantics can significantly degrade the performance of the LM under mismatched conditions. Therefore, this degradation has an adverse effect on the ASR even when dialect-specific AM is used. We show a degradation of up to 13.13 perplexity points when LM is used under mismatched conditions. Furthermore, we show a degradation of over 9% and over 15% in Character Error Rate (CER) and Word Error Rate (WER), respectively, in the ASR systems when using mismatched LMs over matched LMs.

Multimodal large language models for inclusive collaboration learning tasks
Armanda Lewis 15:45-14:15 (Regency A & B)
This PhD project leverages advancements in multimodal large language models to build an inclusive collaboration feedback loop, in order to facilitate the automated detection, modeling, and feedback for participants developing general collaboration skills. This topic is important given the role of collaboration as an essential 21st century skill, the potential to ground large language models within learning theory and real-world practice, and the expressive potential of transformer models to support equity and inclusion. We address some concerns of integrating advances in natural language processing into downstream tasks such as the learning analytics feedback loop.

Industry Oral Session

14:15-15:45 (Quinault)

Self-supervised Product Title Rewrite for Product Listing Ads
Xue Zhao, Dayiheng Liu, Junwei Ding, Liang Yao, Mahone Yan, Huibo Wang and Wenqing Yao 14:15-14:30 (Quinault)
Product Listing Ads (PLAs) are primary online advertisements merchants pay to attract more customers. However, merchants prefer to stack various attributes to the title and neglect the fluency and information priority. These seller-created titles are not suitable for PLAs as they fail to highlight the core information in the visible part in PLAs titles. In this work, we present a title rewrite solution. Specifically, we train a self-supervised language model to generate high-quality titles in terms of fluency and information priority. Extensive offline test and real-world online test have demonstrated that our solution is effective in reducing the cost and gaining more profit as it lowers our CPC, CPB while improving CTR in the online test by a large amount.

Local-to-global learning for iterative training of production SLU models on new features
Yulia Grishina and Daniil Sorokin 14:30-14:45 (Quinault)
In production SLU systems, new training data becomes available with time so that ML models need to be updated on a regular basis. Specifically, releasing new features adds new classes of data while the old data remains constant. However, retraining the full model each time from scratch is computationally expensive. To address this problem, we propose to consider production releases from the curriculum learning perspective and to adapt the local-to-global learning (LGL) schedule (Cheng et. al, 2019) for a statistical model that starts with fewer output classes and adds more classes with each iteration.

We report experiments for the tasks of intent classification and slot filling in the context of a production voice-assistant. First, we apply the original LGL schedule on our data and then adapt LGL to the production setting where the full data is not available at initial training iterations. We demonstrate that our method improves model error rates by 7.3% and saves up to 25% training time for individual iterations.

Medical Coding with Biomedical Transformer Ensembles and Zero/Few-shot Learning
Angelo Ziletti, Alan Akbik, Christoph Berns, Thomas Herold, Marion Legler and Martina Viell 14:45-15:00 (Quinault)
Medical coding (MC) is an essential pre-requisite for reliable data retrieval and reporting. Given a free-text *reported term* (RT) such as “pain of right thigh to the knee”, the task is to identify the matching *lowest-level term* (LLT) –in this case “unilateral leg pain”– from a very large and continuously growing repository of standardized medical terms. However, automating this task is challenging due to a large number of LLT codes (as of writing over 80 000), limited availability of training data for long tail/emerging classes, and the general high accuracy demands of the medical domain. With this paper, we introduce the MC task, discuss its challenges, and present a novel approach called xTARS that combines traditional BERT-based classification with a recent zero/few-shot learning approach (TARS). We present extensive experiments that show that our combined approach outperforms strong baselines, especially in the few-shot regime. The approach is developed and deployed at Bayer, live since November 2021. As we believe our approach potentially promising beyond MC, and to ensure reproducibility, we release the code to the research community.

CTM - A Model for Large-Scale Multi-View Tweet Topic Classification
Vivek Kulkarni, Kenny Leung and Aria Haghighi 15:00-15:15 (Quinault)
Automatically associating social media posts with topics is an important prerequisite for effective search and recommendation on many social media platforms. However, topic classification of such posts is quite challenging because of (a) a large topic space (b) short text with weak topical cues, and (c) multiple topic associations per post. In contrast to most prior work which only focuses on post-classification into a small number of topics ($10-20$), we consider the task of large-scale topic classification in the context of Twitter where the topic space is 10 times larger with potentially multiple topic associations per Tweet. We address the challenges above and propose a novel neural model, that (a) supports a large topic space of 300 topics (b) takes a holistic approach to tweet content modeling – leveraging multi-modal content, author context, and deeper semantic cues in the Tweet. Our method offers an effective way to classify Tweets into topics at scale by yielding superior performance to other approaches (a relative lift of $\mathbf{20\%}$ in median average precision score) and has been successfully deployed in production at Twitter.

Self-Aware Feedback-Based Self-Learning in Large-Scale Conversational AI
Pragaash Ponnusamy, Clint Solomon Mathialagan, Gustavo Aguilar, Chengyuan Ma and Chenlei Guo 15:15-15:30 (Quinault)
Self-learning paradigms in large-scale conversational AI agents tend to leverage user feedback in bridging between what they say and what they mean. However, such learning, particularly in Markov-based query rewriting systems have far from addressed the impact of these models on future training where successive feedback is inevitably contingent on the rewrite itself, especially in a continually updating environment. In this paper, we explore the consequences of this inherent lack of self-awareness towards impairing the model performance, ultimately resulting in both Type I and II errors over time. To that end, we propose augmenting the Markov Graph construction with a superposition-based adjacency matrix. Here, our method leverages an induced stochasticity to reactively learn a locally-adaptive decision boundary based on the

performance of the individual rewrites in a bi-variate beta setting. We also surface a data augmentation strategy that leverages template-based generation in abridging complex conversation hierarchies of dialogs so as to simplify the learning process. All in all, we demonstrate that our self-aware model improves the overall PR-AUC by 27.45%, achieves a relative defect reduction of up to 31.22%, and is able to adapt quicker to changes in global preferences across a large number of customers.

Aspect-based Analysis of Advertising Appeals for Search Engine Advertising
Soichiro Murakami, Peinan Zhang, Sho Hoshino, Hidetaka Kamigaito, Hiroya Takamura and Manabu Okumura 15:30-15:45 (Quinault)
Writing an ad text that attracts people and persuades them to click or act is essential for the success of search engine advertising. Therefore, ad creators must consider various aspects of advertising appeals (A^3) such as the price, product features, and quality. However, products and services exhibit unique effective A^3 for different industries. In this work, we focus on exploring the effective A^3 for different industries with the aim of assisting the ad creation process. To this end, we created a dataset of advertising appeals and used an existing model that detects various aspects for ad texts. Our experiments demonstrated through correlation analysis that different industries have their own effective A^3 and that the identification of the A^3 contributes to the estimation of advertising performance.

Session 6 - 16:15-17:45

Language Grounding to Vision 2

16:15-17:45 (Columbia A)

All You May Need for VQA are Image Captions
Soravit Changpinyo, Doron Kukliansy, Idan Szpektor, Xi Chen, Nan Ding and Radu Soricut 16:15-16:30 (Columbia A)
Visual Question Answering (VQA) has benefited from increasingly sophisticated models, but has not enjoyed the same level of engagement in terms of data creation. In this paper, we propose a method that automatically derives VQA examples at volume, by leveraging the abundance of existing image-caption annotations combined with neural models for textual question generation. We show that the resulting data is of high-quality. VQA models trained on our data improve state-of-the-art zero-shot accuracy by double digits and achieve a level of robustness that lacks in the same model trained on human-annotated VQA data.

Imagination-Augmented Natural Language Understanding
Yujie Lu, Wanrong Zhu, Xin Eric Wang, Miguel Eckstein and William Yang Wang 16:30-16:45 (Columbia A)
Human brains integrate linguistic and perceptual information simultaneously to understand natural language, and hold the critical ability to render imaginations. Such abilities enable us to construct new abstract concepts or concrete objects, and are essential in involving practical knowledge to solve problems in low-resource scenarios. However, most existing methods for Natural Language Understanding (NLU) are mainly focused on textual signals. They do not simulate human visual imagination ability, which hinders models from inferring and learning efficiently from limited data samples. Therefore, we introduce an Imagination-Augmented Cross-modal Encoder (iACE) to solve natural language understanding tasks from a novel learning perspective—imagination-augmented cross-modal understanding. iACE enables visual imagination with external knowledge transferred from the powerful generative and pre-trained vision-and-language models. Extensive experiments on GLUE and SWAG show that iACE achieves consistent improvement over visually-supervised pre-trained models. More importantly, results in extreme and normal few-shot settings validate the effectiveness of iACE in low-resource natural language understanding circumstances.

Visual Commonsense in Pretrained Unimodal and Multimodal Models
Chenyu Zhang, Benjamin Van Durme, Zhuowan Li and Elias Stengel-Eskin 16:45-17:00 (Columbia A)
Our commonsense knowledge about objects includes their typical visual attributes; we know that bananas are typically yellow or green, and not purple. Text and image corpora, being subject to reporting bias, represent this world-knowledge to varying degrees of faithfulness. In this paper, we investigate to what degree unimodal (language-only) and multimodal (image and language) models capture a broad range of visually salient attributes. To that end, we create the Visual Commonsense Tests (ViComTe) dataset covering 5 property types (color, shape, material, size, and visual co-occurrence) for over 5000 subjects. We validate this dataset by showing that our grounded color data correlates much better than ungrounded text-only data with crowdsourced color judgments provided by Paik et al. (2021). We then use our dataset to evaluate pretrained unimodal models and multimodal models. Our results indicate that multimodal models better reconstruct attribute distributions, but are still subject to reporting bias. Moreover, increasing model size does not enhance performance, suggesting that the key to visual commonsense lies in the data.

Few-shot Subgoal Planning with Language Models
Lajanugen Logeswaran, Yao Fu, Moontae Lee and Honglak Lee 17:00-17:15 (Columbia A)
Pre-trained language models have shown successful progress in many text understanding benchmarks. This work explores the capability of these models to predict actionable plans in real-world environments. Given a text instruction, we show that language priors encoded in pre-trained models allow us to infer fine-grained subgoal sequences. In contrast to recent methods which make strong assumptions about subgoal supervision, our experiments show that language models can infer detailed subgoal sequences from few training sequences without any fine-tuning. We further propose a simple strategy to re-rank language model predictions based on interaction and feedback from the environment. Combined with pre-trained navigation and visual reasoning components, our approach demonstrates competitive performance on subgoal prediction and task completion in the ALFRED benchmark compared to prior methods that assume more subgoal supervision.

Disentangling Categorization in Multi-agent Emergent Communication
Washington Garcia, Hamilton Scott Clouse and Kevin R. B. Butler 17:15-17:30 (Columbia A)
The emergence of language between artificial agents is a recent focus of computational linguistics, as it offers a synthetic substrate for reasoning about human language evolution. From the perspective of cognitive science, sophisticated categorization in humans is thought to enable reasoning about novel observations, and thus compose old information to describe new phenomena. Unfortunately, the literature to date has not managed to isolate the effect of categorization power in artificial agents on their inter-communication ability, particularly on novel, unseen objects. In this work, we propose the use of disentangled representations from representation learning to quantify the categorization power of agents, enabling a differential analysis between combinations of heterogeneous systems, e.g., pairs of agents which learn to communicate despite mismatched concept realization. Through this approach, we observe that agent heterogeneity can cut signaling accuracy by up to 40%, despite encouraging compositionality in the artificial language. We conclude that the reasoning process of agents plays a key role in their communication, with unexpected benefits arising from their mixing, such as better language compositionality.

CoSIm: Commonsense Reasoning for Counterfactual Scene Imagination
Hyounghun Kim, Abhay Zala and Mohit Bansal 17:30-17:45 (Columbia A)
As humans, we can modify our assumptions about a scene by imagining alternative objects or concepts in our minds. For example, we can easily anticipate the implications of the sun being overcast by rain clouds (e.g., the street will get wet) and accordingly prepare for that. In this paper, we introduce a new dataset called Commonsense Reasoning for Counterfactual Scene Imagination (CoSIm) which is designed to evaluate the ability of AI systems to reason about scene change imagination. To be specific, in this multimodal task/dataset, models are given an image and an initial question-response pair about the image. Next, a counterfactual imagined scene change (in textual form) is applied, and the model has to predict the new response to the initial question based on this scene change. We collect 3.5K high-quality and challenging data instances, with each instance consisting of an image, a commonsense question with a response, a description of a counterfactual change, a new response to the question, and three distractor responses. Our dataset contains various complex scene change types (such as object addition/removal/state change, event description, environment change, etc.) that require models to imagine many different scenarios and reason about the changed scenes. We present a baseline model based on a vision-language Transformer (i.e., LXMERT) and ablation studies. Through human evaluation, we demonstrate a large human-model performance gap, suggesting room for promising future work on this challenging, counterfactual multimodal task.

Syntax: Tagging, Chunking and Parsing

16:15-17:45 (Columbia C)

[CL] **Tractable Parsing for CCGs of Bounded Degree**
Lena Katharina Schiffer, Marco Kuhlmann and Giorgio Satta 16:15-16:30 (Columbia C)
Unlike other mildly context-sensitive formalisms, Combinatory Categorial Grammar (CCG) cannot be parsed in polynomial time when the size of the grammar is taken into account. Refining this result, we show that the parsing complexity of CCG is exponential only in the maximum degree of composition. When that degree is fixed, parsing can be carried out in polynomial time. Our finding is interesting from a linguistic perspective because a bounded degree of composition has been suggested as a universal constraint on natural language grammar. Moreover, ours is the first complexity result for a version of CCG that includes substitution rules, which are used in practical grammars but have been ignored in theoretical work

Template-free Prompt Tuning for Few-shot NER
Ruotian Ma, Xin Zhou, Tao Gui, Yiding Tan, Linyang Li, Qi Zhang and Xuanjing Huang 16:30-16:45 (Columbia C)
Prompt-based methods have been successfully applied in sentence-level few-shot learning tasks, mostly owing to the sophisticated design of templates and label words. However, when applied to token-level labeling tasks such as NER, it would be time-consuming to enumerate the template queries over all potential entity spans. In this work, we propose a more elegant method to reformulate NER tasks as LM problems without any templates. Specifically, we discard the template construction process while maintaining the word prediction paradigm of pre-training models to predict a class-related pivot word (or label word) at the entity position. Meanwhile, we also explore principled ways to automatically search for appropriate label words that the pre-trained models can easily adapt to. While avoiding the complicated template-based process, the proposed LM objective also reduces the gap between different objectives used in pre-training and fine-tuning, thus it can better benefit the few-shot performance. Experimental results demonstrate the effectiveness of the proposed method over bert-tagger and template-based method under few-shot settings. Moreover, the decoding speed of the proposed method is up to 1930.12 times faster than the template-based method.

Dynamic Gazetteer Integration in Multilingual Models for Cross-Lingual and Cross-Domain Named Entity Recognition
Besnik Fetahu, Anjie Fang, Oleg Rokhlenko and Shervin Malmasi 16:45-17:00 (Columbia C)
Named entity recognition (NER) in a real-world setting remains challenging and is impacted by factors like text genre, corpus quality, and data availability. NER models trained on CoNLL do not transfer well to other domains, even within the same language. This is especially the case for multi-lingual models when applied to low-resource languages, and is mainly due to missing entity information.

We propose an approach that with limited effort and data, addresses the NER knowledge gap across languages and domains. Our novel approach uses a token-level gating layer to augment pre-trained multilingual transformers with gazetteers containing named entities (NE) from a target language or domain. This approach provides the flexibility to jointly integrate both textual and gazetteer information dynamically: entity knowledge from gazetteers is used only when a token's textual representation is insufficient for the NER task.

Evaluation on several languages and domains demonstrates: (i) a high mismatch of reported NER performance on CoNLL vs. domain specific datasets, (ii) gazetteers significantly improve NER performance across languages and domains, and (iii) gazetteers can be flexibly incorporated to guide knowledge transfer. On cross-lingual transfer we achieve an improvement over the baseline with F1=+17.6%, and with F1=+21.3% for cross-domain transfer.

Unsupervised Cross-Lingual Transfer of Structured Predictors without Source Data
Kemal Kurniawan, Lea Frermann, Philip Schulz and Trevor Cohn 17:00-17:15 (Columbia C)
Providing technologies to communities or domains where training data is scarce or protected e.g., for privacy reasons, is becoming increasingly important. To that end, we generalise methods for unsupervised transfer from multiple input models for structured prediction. We show that the means of aggregating over the input models is critical, and that multiplying marginal probabilities of substructures to obtain high-probability structures for distant supervision is substantially better than taking the union of such structures over the input models, as done in prior work. Testing on 18 languages, we demonstrate that the method works in a cross-lingual setting, considering both dependency parsing and part-of-speech structured prediction problems. Our analyses show that the proposed method produces less noisy labels for the distant supervision.

Masked Part-Of-Speech Model: Does Modeling Long Context Help Unsupervised POS-tagging?
Xiang Zhou, Shiyue Zhang and Mohit Bansal 17:15-17:30 (Columbia C)
Previous Part-Of-Speech (POS) induction models usually assume certain independence assumptions (e.g., Markov, unidirectional, local dependency) that do not hold in real languages. For example, the subject-verb agreement can be both long-term and bidirectional. To facilitate flexible dependency modeling, we propose a Masked Part-of-Speech Model (MPoSM), inspired by the recent success of Masked Language Models (MLM). MPoSM can model arbitrary tag dependency and perform POS induction through the objective of masked POS reconstruction. We achieve competitive results on both the English Penn WSJ dataset as well as the universal treebank containing 10 diverse languages. Though modeling the long-term dependency should ideally help this task, our ablation study shows mixed trends in different languages. To better understand this phenomenon, we design a novel synthetic experiment that can specifically diagnose the model's ability to learn tag agreement. Surprisingly, we find that even strong baselines fail to solve this problem consistently in a very simplified setting: the agreement

between adjacent words. Nonetheless, MPoSM achieves overall better performance. Lastly, we conduct a detailed error analysis to shed light on other remaining challenges.

Dynamic Programming in Rank Space: Scaling Structured Inference with Low-Rank HMMs and PCFGs
Songlin Yang, Wei Liu and Kewei Tu 17:30-17:45 (Columbia C)
Hidden Markov Models (HMMs) and Probabilistic Context-Free Grammars (PCFGs) are widely used structured models, both of which can be represented as factor graph grammars (FGGs), a powerful formalism capable of describing a wide range of models. Recent research found it beneficial to use large state spaces for HMMs and PCFGs. However, inference with large state spaces is computationally demanding, especially for PCFGs. To tackle this challenge, we leverage tensor rank decomposition (aka. CPD) to decrease inference computational complexities for a subset of FGGs subsuming HMMs and PCFGs. We apply CPD on the factors of an FGG and then construct a new FGG defined in the rank space. Inference with the new FGG produces the same result but has a lower time complexity when the rank size is smaller than the state size. We conduct experiments on HMM language modeling and unsupervised PCFG parsing, showing better performance than previous work. Our code is publicly available at `https://github.com/VPeterV/RankSpace-Models`.

Multilinguality

16:15-17:45 (Columbia D)

A Balanced Data Approach for Evaluating Cross-Lingual Transfer: Mapping the Linguistic Blood Bank
Dan Malkin, Tomasz Limisiewicz and Gabriel Stanovsky 16:15-16:30 (Columbia D)
We show that the choice of pretraining languages affects downstream cross-lingual transfer for BERT-based models. We inspect zero-shot performance in balanced data conditions to mitigate data size confounds, classifying pretraining languages that improve downstream performance as donors, and languages that are improved in zero-shot performance as recipients. We develop a method of quadratic time complexity in the number of languages to estimate these relations, instead of an exponential exhaustive computation of all possible combinations. We find that our method is effective on a diverse set of languages spanning different linguistic features and two downstream tasks. Our findings can inform developers of large-scale multilingual language models in choosing better pretraining configurations.

When is BERT Multilingual? Isolating Crucial Ingredients for Cross-lingual Transfer
Ameet Deshpande, Partha Talukdar and Karthik R Narasimhan 16:30-16:45 (Columbia D)
While recent work on multilingual language models has demonstrated their capacity for cross-lingual zero-shot transfer on downstream tasks, there is a lack of consensus in the community as to what shared properties between languages enable such transfer. Analyses involving pairs of natural languages are often inconclusive and contradictory since languages simultaneously differ in many linguistic aspects. In this paper, we perform a large-scale empirical study to isolate the effects of various linguistic properties by measuring zero-shot transfer between four diverse natural languages and their counterparts constructed by modifying aspects such as the script, word order, and syntax. Among other things, our experiments show that the absence of sub-word overlap significantly affects zero-shot transfer when languages differ in their word order, and there is a strong correlation between transfer performance and word embedding alignment between languages (e.g., $\rho\_s = 0.94$ on the task of NLI). Our results call for focus in multilingual models on explicitly improving word embedding alignment between languages rather than relying on its implicit emergence.

Lifting the Curse of Multilinguality by Pre-training Modular Transformers
Jonas Pfeiffer, Naman Goyal, Xi Victoria Lin, Xian Li, James Cross, Sebastian Riedel and Mikel Artetxe 16:45-17:00 (Columbia D)
Multilingual pre-trained models are known to suffer from the curse of multilinguality, which causes per-language performance to drop as they cover more languages. We address this issue by introducing language-specific modules, which allows us to grow the total capacity of the model, while keeping the total number of trainable parameters per language constant. In contrast with prior work that learns language-specific components post-hoc, we pre-train the modules of our Cross-lingual Modular (X-Mod) models from the start. Our experiments on natural language inference, named entity recognition and question answering show that our approach not only mitigates the negative interference between languages, but also enables positive transfer, resulting in improved monolingual and cross-lingual performance. Furthermore, our approach enables adding languages post-hoc with no measurable drop in performance, no longer limiting the model usage to the set of pre-trained languages.

Combating the Curse of Multilinguality in Cross-Lingual WSD by Aligning Sparse Contextualized Word Representations
Gábor Berend 17:00-17:15 (Columbia D)
In this paper, we advocate for using large pre-trained monolingual language models in cross lingual zero-shot word sense disambiguation (WSD) coupled with a contextualized mapping mechanism. We also report rigorous experiments that illustrate the effectiveness of employing sparse contextualized word representations obtained via a dictionary learning procedure. Our experimental results demonstrate that the above modifications yield a significant improvement of nearly 6.5 points of increase in the average F-score (from 62.0 to 68.5) over a collection of 17 typologically diverse set of target languages. We release our source code for replicating our experiments at https://github.com/begab/sparsity_makes_sense.

On the Economics of Multilingual Few-shot Learning: Modeling the Cost-Performance Trade-offs of Machine Translated and Manual Data
Kabir Ahuja, Monojit Choudhury and Sandipan Dandapat 17:15-17:30 (Columbia D)
Borrowing ideas from Production functions in micro-economics, in this paper we introduce a framework to systematically evaluate the performance and cost trade-offs between machine-translated and manually-created labelled data for task-specific fine-tuning of massively multilingual language models. We illustrate the effectiveness of our framework through a case-study on the TyDIQA-GoldP dataset. One of the interesting conclusion of the study is that if the cost of machine translation is greater than zero, the optimal performance at least cost is always achieved with at least some or only manually-created data. To our knowledge, this is the first attempt towards extending the concept of production functions to study data collection strategies for training multilingual models, and can serve as a valuable tool for other similar cost vs data trade-offs in NLP.

Bridging the Gap between Language Models and Cross-Lingual Sequence Labeling
Nuo Chen, Linjun Shou, Ming Gong, Jian Pei and Daxin Jiang 17:30-17:45 (Columbia D)
Large-scale cross-lingual pre-trained language models (xPLMs) have shown effective in cross-lingual sequence labeling tasks (xSL), such as machine reading comprehension (xMRC) by transferring knowledge from a high-resource language to low-resource languages. Despite the great success, we draw an empirical observation that there is an training objective gap between pre-training and fine-tuning stages: e.g., mask language modeling objective requires *local* understanding of the masked token and the span-extraction objective requires understanding and

reasoning of the *global* input passage/paragraph and question, leading to the discrepancy between pre-training and xMRC. In this paper, we first design a pre-training task tailored for xSL named Cross-lingual Language Informative Span Masking (CLISM) to eliminate the objective gap in a self-supervised manner. Second, we present ContrAstive-Consistency Regularization (CACR), which utilizes contrastive learning to encourage the consistency between representations of input parallel sequences via unsupervised cross-lingual instance-wise training signals during pre-training. By these means, our methods not only bridge the gap between pretrain-finetune, but also enhance PLMs to better capture the alignment between different languages. Extensive experiments prove that our method achieves clearly superior results on multiple xSL benchmarks with limited pre-training data. Our methods also surpass the previous state-of-the-art methods by a large margin in few-shot data setting, where only a few hundred training examples are available.

Machine Learning for NLP 1

16:15-17:45 (Elwha A)

DEMix Layers: Disentangling Domains for Modular Language Modeling
Suchin Gururangan, Mike Lewis, Ari Holtzman, Noah Smith and Luke Zettlemoyer 16:15-16:30 (Elwha A)
We introduce a new domain expert mixture (DEMix) layer that enables conditioning a language model (LM) on the domain of the input text. A DEMix layer includes a collection of expert feedforward networks, each specialized to a domain, that makes the LM modular: experts can be mixed, added, or removed after initial training. Extensive experiments with autoregressive transformer LMs (up to 1.3B parameters) show that DEMix layers reduce test-time perplexity (especially for out-of-domain data), increase training efficiency, and enable rapid adaptation. Mixing experts during inference, using a parameter-free weighted ensemble, enables better generalization to heterogeneous or unseen domains. We also show it is possible to add experts to adapt to new domains without forgetting older ones, and remove experts to restrict access to unwanted domains. Overall, these results demonstrate benefits of domain modularity in language models.

Practice Makes a Solver Perfect: Data Augmentation for Math Word Problem Solvers
Vivek Kumar, Rishabh Maheshwary and Vikram Pudi 16:30-16:45 (Elwha A)
Existing Math Word Problem (MWP) solvers have achieved high accuracy on benchmark datasets. However, prior works have shown that such solvers do not generalize well and rely on superficial cues to achieve high performance. In this paper, we first conduct experiments to showcase that this behaviour is mainly associated with the limited size and diversity present in existing MWP datasets. Next, we propose several data augmentation techniques broadly categorized into Substitution and Paraphrasing based methods. By deploying these methods we increase the size of existing datasets by five folds. Extensive experiments on two benchmark datasets across three state-of-the-art MWP solvers shows that proposed methods increase the generalization and robustness of existing solvers. On average, proposed methods significantly increase the state-of-the-art results by over five percentage points on benchmark datasets. Further, the solvers trained on the augmented dataset performs comparatively better on the challenge test set. We also show the effectiveness of proposed techniques through ablation studies and verify the quality of augmented samples through human evaluation.

Quantifying Adaptability in Pre-trained Language Models with 500 Tasks
Belinda Z. Li, Jane A. Yu, Madian Khabsa, Luke Zettlemoyer, Alon Y. Halevy and Jacob Andreas 16:45-17:00 (Elwha A)
When a neural language model (LM) is adapted to perform a new task, what aspects of the task predict the eventual performance of the model? In NLP, systematic features of LM generalization to individual examples are well characterized, but systematic aspects of LM adaptability to new tasks are not nearly as well understood. We present a large-scale empirical study of the features and limits of LM adaptability using a new benchmark, TaskBench500, built from 500 procedurally generated sequence modeling tasks. These tasks combine core aspects of language processing, including lexical semantics, sequence processing, memorization, logical reasoning, and world knowledge. Using TaskBench500, we evaluate three facets of adaptability, finding that: (1) adaptation procedures differ dramatically in their ability to memorize small datasets; (2) within a subset of task types, adaptation procedures exhibit compositional adaptability to complex tasks; and (3) failure to match training label distributions is explained by mismatches in the intrinsic difficulty of predicting individual labels. Our experiments show that adaptability to new tasks, like generalization to new examples, can be systematically described and understood, and we conclude with a discussion of additional aspects of adaptability that could be studied using the new benchmark.

KALA: Knowledge-Augmented Language Model Adaptation
Minki Kang, Jinheon Baek and Sung Ju Hwang 17:00-17:15 (Elwha A)
Pre-trained language models (PLMs) have achieved remarkable success on various natural language understanding tasks. Simple fine-tuning of PLMs, on the other hand, might be suboptimal for domain-specific tasks because they cannot possibly cover knowledge from all domains. While adaptive pre-training of PLMs can help them obtain domain-specific knowledge, it requires a large training cost. Moreover, adaptive pre-training can harm the PLM's performance on the downstream task by causing catastrophic forgetting of its general knowledge. To overcome such limitations of adaptive pre-training for PLM adaption, we propose a novel domain adaption framework for PLMs coined as Knowledge-Augmented Language model Adaptation (KALA), which modulates the intermediate hidden representations of PLMs with domain knowledge, consisting of entities and their relational facts. We validate the performance of our KALA on question answering and named entity recognition tasks on multiple datasets across various domains. The results show that, despite being computationally efficient, our KALA largely outperforms adaptive pre-training.

Extreme Zero-Shot Learning for Extreme Text Classification
Yuanhao Xiong, Wei-Cheng Chang, Cho-Jui Hsieh, Hsiang-Fu Yu and Inderjit S Dhillon 17:15-17:30 (Elwha A)
The eXtreme Multi-label text Classification (XMC) problem concerns finding most relevant labels for an input text instance from a large label set. However, the XMC setup faces two challenges: (1) it is not generalizable to predict unseen labels in dynamic environments, and (2) it requires a large amount of supervised (instance, label) pairs, which can be difficult to obtain for emerging domains. In this paper, we consider a more practical scenario called Extreme Zero-Shot XMC (EZ-XMC), in which no supervision is needed and merely raw text of instances and labels are accessible. Few-Shot XMC (FS-XMC), an extension to EZ-XMC with limited supervision is also investigated. To learn the semantic embeddings of instances and labels with raw text, we propose to pre-train Transformer-based encoders with self-supervised contrastive losses. Specifically, we develop a pre-training method **MACLR**, which thoroughly leverages the raw text with techniques including **M**ulti-scale **A**daptive **C**lustering, **L**abel **R**egularization, and self-training with pseudo positive pairs. Experimental results on four public EZ-XMC datasets demonstrate that MACLR achieves superior performance compared to all other leading baseline methods, in particular with approximately 5-10% improvement in precision and recall on average. Moreover, we show that our pre-trained encoder can be further improved on FS-XMC when there are a limited number of ground-truth positive pairs in training. Our code is available at https://github.com/amzn/pecos/tree/mainline/examples/MACLR.

TreeMix: Compositional Constituency-based Data Augmentation for Natural Language Understanding

Le Zhang, Zichao Yang and Diyi Yang 17:30-17:45 (Elwha A)
Data augmentation is an effective approach to tackle over-fitting. Many previous works have proposed different data augmentations strategies for NLP, such as noise injection, word replacement, back-translation etc. Though effective, they missed one important characteristic of language–compositionality, meaning of a complex expression is built from its sub-parts. Motivated by this, we propose a compositional data augmentation approach for natural language understanding called TreeMix. Specifically, TreeMix leverages constituency parsing tree to decompose sentences into constituent sub-structures and the Mixup data augmentation technique to recombine them to generate new sentences. Compared with previous approaches, TreeMix introduces greater diversity to the samples generated and encourages models to learn compositionality of NLP data. Extensive experiments on text classification and SCAN demonstrate that TreeMix outperforms current state-of-the-art data augmentation methods.

Question Answering 1

16:15-17:45 (Elwha B)

QuALITY: Question Answering with Long Input Texts, Yes!
Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny L Ma, Jana Thompson, He He and Samuel R. Bowman 16:15-16:30 (Elwha B)
To enable building and testing models on long-document comprehension, we introduce QuALITY, a multiple-choice QA dataset with context passages in English that have an average length of about 5,000 tokens, much longer than typical current models can process. Unlike in prior work with passages, our questions are written and validated by contributors who have read the entire passage, rather than relying on summaries or excerpts. In addition, only half of the questions are answerable by annotators working under tight time constraints, indicating that skimming and simple search are not enough to consistently perform well. Our baseline models perform poorly on this task (55.4%) and significantly lag behind human performance (93.5%).

On the Robustness of Reading Comprehension Models to Entity Renaming
Jun Yan, Yang Xiao, Sagnik Mukherjee, Bill Yuchen Lin, Robin Jia and Xiang Ren 16:30-16:45 (Elwha B)
We study the robustness of machine reading comprehension (MRC) models to entity renaming—do models make more wrong predictions when the same questions are asked about an entity whose name has been changed? Such failures imply that models overly rely on entity information to answer questions, and thus may generalize poorly when facts about the world change or questions are asked about novel entities. To systematically audit this issue, we present a pipeline to automatically generate test examples at scale, by replacing entity names in the original test sample with names from a variety of sources, ranging from names in the same test set, to common names in life, to arbitrary strings. Across five datasets and three pretrained model architectures, MRC models consistently perform worse when entities are renamed, with particularly large accuracy drops on datasets constructed via distant supervision. We also find large differences between models: SpanBERT, which is pretrained with span-level masking, is more robust than RoBERTa, despite having similar accuracy on unperturbed test data. We further experiment with different masking strategies as the continual pretraining objective and find that entity-based masking can improve the robustness of MRC models.

OmniTab: Pretraining with Natural and Synthetic Data for Few-shot Table-based Question Answering
Zhengbao Jiang, Yi Mao, Pengcheng He, Graham Neubig and Weizhu Chen 16:45-17:00 (Elwha B)
The information in tables can be an important complement to text, making table-based question answering (QA) systems of great value. The intrinsic complexity of handling tables often adds an extra burden to both model design and data annotation. In this paper, we aim to develop a simple table-based QA model with minimal annotation effort. Motivated by the fact that table-based QA requires both alignment between questions and tables and the ability to perform complicated reasoning over multiple table elements, we propose an omnivorous pretraining approach that consumes both natural and synthetic data to endow models with these respective abilities. Specifically, given freely available tables, we leverage retrieval to pair them with relevant natural sentences for mask-based pretraining, and synthesize NL questions by converting SQL sampled from tables for pretraining with a QA loss. We perform extensive experiments in both few-shot and full settings, and the results clearly demonstrate the superiority of our model OmniTab, with the best multitasking approach achieving an absolute gain of 16.2% and 2.7% in 128-shot and full settings respectively, also establishing a new state-of-the-art on WikiTableQuestions. Detailed ablations and analyses reveal different characteristics of natural and synthetic data, shedding light on future directions in omnivorous pretraining.

Modeling Exemplification in Long-form Question Answering via Retrieval
Shufan Wang, Fangyuan Xu, Laure Thompson, Eunsol Choi and Mohit Iyyer 17:00-17:15 (Elwha B)
Exemplification is a process by which writers explain or clarify a concept by providing an example. While common in all forms of writing, exemplification is particularly useful in the task of long-form question answering (LFQA), where a complicated answer can be made more understandable through simple examples. In this paper, we provide the first computational study of exemplification in QA, performing a fine-grained annotation of different types of examples (e.g., hypotheticals, anecdotes) in three corpora. We show that not only do state-of-the-art LFQA models struggle to generate relevant examples, but also that standard evaluation metrics such as ROUGE are insufficient to judge exemplification quality. We propose to treat exemplification as a *retrieval* problem in which a partially-written answer is used to query a large set of human-written examples extracted from a corpus. Our approach allows a reliable ranking-type automatic metrics that correlates well with human evaluation. A human evaluation shows that our model's retrieved examples are more relevant than examples generated from a state-of-the-art LFQA model.

JointLK: Joint Reasoning with Language Models and Knowledge Graphs for Commonsense Question Answering
Yueqing Sun, Qi Shi, Le Qi and Yu Zhang 17:15-17:30 (Elwha B)
Existing KG-augmented models for commonsense question answering primarily focus on designing elaborate Graph Neural Networks (GNNs) to model knowledge graphs (KGs). However, they ignore (i) the effectively fusing and reasoning over question context representations and the KG representations, and (ii) automatically selecting relevant nodes from the noisy KGs during reasoning. In this paper, we propose a novel model, JointLK, which solves the above limitations through the joint reasoning of LM and GNN and the dynamic KGs pruning mechanism. Specifically, JointLK performs joint reasoning between LM and GNN through a novel dense bidirectional attention module, in which each question token attends on KG nodes and each KG node attends on question tokens, and the two modal representations fuse and update mutually by multi-step interactions. Then, the dynamic pruning module uses the attention weights generated by joint reasoning to prune irrelevant KG nodes recursively. We evaluate JointLK on the CommonsenseQA and OpenBookQA datasets, and demonstrate its improvements to the existing LM and LM+KG models, as well as its capability to perform interpretable reasoning.

Clues Before Answers: Generation-Enhanced Multiple-Choice QA
Zixian Huang, Ao Wu, Jiaying Zhou, Yu Gu, Yue Zhao and Gong Cheng 17:30-17:45 (Elwha B)

A trending paradigm for multiple-choice question answering (MCQA) is using a text-to-text framework. By unifying data in different tasks into a single text-to-text format, it trains a generative encoder-decoder model which is both powerful and universal. However, a side effect of twisting a generation target to fit the classification nature of MCQA is the under-utilization of the decoder and the knowledge that can be decoded. To exploit the generation capability and underlying knowledge of a pre-trained encoder-decoder model, in this paper, we propose a generation-enhanced MCQA model named GenMC. It generates a clue from the question and then leverages the clue to enhance a reader for MCQA. It outperforms text-to-text models on multiple MCQA datasets.

Virtual Poster Q&A Session 2

16:15-17:45 (702 Clearwater)

Knowledge-Grounded Dialogue Generation with a Unified Knowledge Representation
Yu Li, Baolin Peng, Yelong Shen, Yi Mao, Lars Liden, Zhou Yu and Jianfeng Gao 16:15-17:45 (702 Clearwater)
Knowledge-grounded dialogue systems are challenging to build due to the lack of training data and heterogeneous knowledge sources. Existing systems perform poorly on unseen topics due to limited topics covered in the training data. In addition, it is challenging to generalize to the domains that require different types of knowledge sources. To address the above challenges, we present PLUG, a language model that homogenizes different knowledge sources to a unified knowledge representation for knowledge-grounded dialogue generation tasks. We first retrieve relevant information from heterogeneous knowledge sources (e.g., wiki, dictionary, or knowledge graph); Then the retrieved knowledge is transformed into text and concatenated with dialogue history to feed into the language model for generating responses. PLUG is pre-trained on a large-scale knowledge-grounded dialogue corpus. The empirical evaluation on two benchmarks shows that PLUG generalizes well across different knowledge-grounded dialogue tasks. It achieves comparable performance with state-of-the-art methods in the fully-supervised setting and significantly outperforms other approaches in zero-shot and few-shot settings.

Learning Dialogue Representations from Consecutive Utterances
Zhihan Zhou, Dejiao Zhang, Wei Xiao, Nicholas Dingwall, Xiaofei Ma, Andrew Arnold and Bing Xiang 16:15-17:45 (702 Clearwater)
Learning high-quality dialogue representations is essential for solving a variety of dialogue-oriented tasks, especially considering that dialogue systems often suffer from data scarcity. In this paper, we introduce Dialogue Sentence Embedding (DSE), a self-supervised contrastive learning method that learns effective dialogue representations suitable for a wide range of dialogue tasks. DSE learns from dialogues by taking consecutive utterances of the same dialogue as positive pairs for contrastive learning. Despite its simplicity, DSE achieves significantly better representation capability than other dialogue representation and universal sentence representation models. We evaluate DSE on five downstream dialogue tasks that examine dialogue representation at different semantic granularities. Experiments in few-shot and zero-shot settings show that DSE outperforms baselines by a large margin, for example, it achieves 13% average performance improvement over the strongest unsupervised baseline in 1-shot intent classification on 6 datasets. We also provide analyses on the benefits and limitations of our model.

Emp-RFT: Empathetic Response Generation via Recognizing Feature Transitions between Utterances
Wongyu Kim, Youbin Ahn, Donghyun Kim and Kyong-Ho Lee 16:15-17:45 (702 Clearwater)
Each utterance in multi-turn empathetic dialogues has features such as emotion, keywords, and utterance-level meaning. Feature transitions between utterances occur naturally. However, existing approaches fail to perceive the transitions because they extract features for the context at the coarse-grained level. To solve the above issue, we propose a novel approach of recognizing feature transitions between utterances, which helps understand the dialogue flow and better grasp the features of utterance that needs attention. Also, we introduce a response generation strategy to help focus on emotion and keywords related to appropriate features when generating responses. Experimental results show that our approach outperforms baselines and especially, achieves significant improvements on multi-turn dialogues.

Show, Don't Tell: Demonstrations Outperform Descriptions for Schema-Guided Task-Oriented Dialogue
Raghav Gupta, Harrison Lee, Jeffrey Zhao, Yuan Cao, Abhinav Rastogi and Yonghui Wu 16:15-17:45 (702 Clearwater)
Building universal dialogue systems that operate across multiple domains/APIs and generalize to new ones with minimal overhead is a critical challenge. Recent works have leveraged natural language descriptions of schema elements to enable such systems; however, descriptions only indirectly convey schema semantics. In this work, we propose Show, Don't Tell, which prompts seq2seq models with a labeled example dialogue to show the semantics of schema elements rather than tell the model through descriptions. While requiring similar effort from service developers as generating descriptions, we show that using short examples as schema representations with large language models results in state-of-the-art performance on two popular dialogue state tracking benchmarks designed to measure zero-shot generalization - the Schema-Guided Dialogue dataset and the MultiWOZ leave-one-out benchmark.

Disentangling Indirect Answers to Yes-No Questions in Real Conversations
Krishna Chaitanya Sanagavarapu, Jathin Pranav Singaraju, Anusha Kakileti, Anirudh Kaza, Aaron Abraham Mathews, Helen Li, Nathan Raul Brito and Eduardo Blanco 16:15-17:45 (702 Clearwater)
In this paper, we explore the task of determining indirect answers to yes-no questions in real conversations. We work with transcripts of phone conversations in the Switchboard Dialog Act (SwDA) corpus and create SwDA-IndirectAnswers (SwDA-IA), a subset of SwDA consisting of all conversations containing a yes-no question with an indirect answer. We annotate the underlying direct answers to the yes-no questions (yes, probably yes, middle, probably no, or no). We show that doing so requires taking into account conversation context: the indirect answer alone is insufficient to determine the ground truth. Experimental results also show that taking into account context is beneficial. More importantly, our results demonstrate that existing corpora with synthetic indirect answers to yes-no questions are not beneficial when working with real conversations. Our best models outperform the majority baseline by a substantial margin, but the task remains a challenge (F1: 0.46).

On the Origin of Hallucinations in Conversational Models: Is it the Datasets or the Models?
Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane and Siva Reddy 16:15-17:45 (702 Clearwater)
Knowledge-grounded conversational models are known to suffer from producing factually invalid statements, a phenomenon commonly called hallucination. In this work, we investigate the underlying causes of this phenomenon: is hallucination due to the training data, or to the models? We conduct a comprehensive human study on both existing knowledge-grounded conversational benchmarks and several state-of-the-art models. Our study reveals that the standard benchmarks consist of > 60% hallucinated responses, leading to models that not only hallucinate but even amplify hallucinations. Our findings raise important questions on the quality of existing datasets and models trained using them. We make our annotations publicly available for future research.

Instilling Type Knowledge in Language Models via Multi-Task QA
Shuyang Li, Mukund Sridhar, Chandana Satya Prakash, Jin Cao, Wael Hamza and Julian McAuley 16:15-17:45 (702 Clearwater)

Understanding human language often necessitates understanding entities and their place in a taxonomy of knowledge—their *types*. Previous methods to learn entity types rely on training classifiers on datasets with coarse, noisy, and incomplete labels. We introduce a method to instill fine-grained type knowledge in language models with text-to-text pre-training on type-centric questions leveraging knowledge base documents and knowledge graphs. We create the **WikiWiki** dataset: entities and passages from 10M Wikipedia articles linked to the Wikidata knowledge graph with 41K types. Models trained on WikiWiki achieve state-of-the-art performance in zero-shot dialog state tracking benchmarks, accurately infer entity types in Wikipedia articles, and can discover new types deemed useful by human judges.

A Versatile Adaptive Curriculum Learning Framework for Task-oriented Dialogue Policy Learning
Yang Yang Zhao, Hua Qin, Wang Zhenyu, Changxi Zhu and Shihan Wang 16:15-17:45 (702 Clearwater)
Training a deep reinforcement learning-based dialogue policy with brute-force random sampling is costly. A new training paradigm was proposed to improve learning performance and efficiency by combining curriculum learning. However, attempts in the field of dialogue policy are very limited due to the lack of reliable evaluation of difficulty scores of dialogue tasks and the high sensitivity to the mode of progression through dialogue tasks. In this paper, we present a novel versatile adaptive curriculum learning (VACL) framework, which presents a substantial step toward applying automatic curriculum learning on dialogue policy tasks. It supports evaluating the difficulty of dialogue tasks only using the learning experiences of dialogue policy and skip-level selection according to their learning needs to maximize the learning efficiency. Moreover, an attractive feature of VACL is the construction of a generic, elastic global curriculum while training a good dialogue policy that could guide different dialogue policy learning without extra effort on re-training. The superiority and versatility of VACL are validated on three public dialogue datasets.

Go Back in Time: Generating Flashbacks in Stories with Event Temporal Prompts
Rujun Han, Hong Chen, Yufei Tian and Nanyun Peng 16:15-17:45 (702 Clearwater)
Stories or narratives are comprised of a sequence of events. To compose interesting stories, professional writers often leverage a creative writing technique called *flashback* that inserts past events into current storylines as we commonly observe in novels and plays. However, it is challenging for machines to generate *flashback* as it requires a solid understanding of event **temporal order** (e.g. *feeling hungry* before *eat*, not vice versa), and the creativity to arrange storylines so that earlier events do not always appear first in **narrative order**. Two major issues in existing systems that exacerbate the challenges: 1) temporal bias in pertaining and story datasets that leads to monotonic event temporal orders; 2) lack of explicit guidance that helps machines decide where to insert *flashbacks*. We propose to address these issues using structured storylines to encode events and their pair-wise temporal relations (before, after and vague) as **temporal prompts** that guide how stories should unfold temporally. We leverage a Plan-and-Write framework enhanced by reinforcement learning to generate storylines and stories end-to-end. Evaluation results show that the proposed method can generate more interesting stories with *flashbacks* while maintaining textual diversity, fluency, and temporal coherence.

Syntax Controlled Knowledge Graph-to-Text Generation with Order and Semantic Consistency
Jin Liu, Chongfeng Fan, Zhou Fengyu and Huijuan Xu 16:15-17:45 (702 Clearwater)
The knowledge graph (KG) stores a large amount of structural knowledge, while it is not easy for direct human understanding. Knowledge graph-to-text (KG-to-text) generation aims to generate easy-to-understand sentences from the KG, and at the same time, maintains semantic consistency between generated sentences and the KG. Existing KG-to-text generation methods phrase this task as a sequence-to-sequence generation task with linearized KG as input and consider the consistency issue of the generated texts and KG through a simple selection between decoded sentence word and KG node word at each time step. However, the linearized KG order is obtained through a heuristic search without data-driven optimization. In this paper, we optimize the knowledge description order prediction under the order supervision extracted from the caption and further enhance the consistency of the generated sentences and KG through syntactic and semantic regularization. We incorporate the Part-of-Speech (POS) syntactic tags to constrain the positions to copy words from the KG and employ a semantic context scoring function to evaluate the semantic fitness for each word in its local context when decoding each word in the generated sentence. Extensive experiments are conducted on two datasets, WebNLG and DART, and achieve state-of-the-art performances. Our code is now public available[3].

TSTR: Too Short to Represent, Summarize with Details! Intro-Guided Extended Summary Generation
Sajad Sotudeh and Nazli Goharian 16:15-17:45 (702 Clearwater)
Many scientific papers such as those in arXiv and PubMed data collections have abstracts with varying lengths of 50-1000 words and average length of approximately 200 words, where longer abstracts typically convey more information about the source paper. Up to recently, scientific summarization research has typically focused on generating short, abstract-like summaries following the existing datasets used for scientific summarization. In domains where the source text is relatively long-form, such as in scientific documents, such summary is not able to go beyond the general and coarse overview and provide salient information from the source document. The recent interest to tackle this problem motivated curation of scientific datasets, arXiv-Long and PubMed-Long, containing human-written summaries of 400-600 words, hence, providing a venue for research in generating long/extended summaries. Extended summaries facilitate a faster read while providing details beyond coarse information. In this paper, we propose TSTR, an extractive summarizer that utilizes the introductory information of documents as pointers to their salient information. The evaluations on two existing large-scale extended summarization datasets indicate statistically significant improvement in terms of Rouge and average Rouge (F1) scores (except in one case) as compared to strong baselines and state-of-the-art. Comprehensive human evaluations favor our generated extended summaries in terms of cohesion and completeness.

Improving Multi-Document Summarization through Referenced Flexible Extraction with Credit-Awareness
Yun-Zhu Song, Yi-Syuan Chen and Hong-Han Shuai 16:15-17:45 (702 Clearwater)
A notable challenge in Multi-Document Summarization (MDS) is the extremely-long length of the input. In this paper, we present an extract-then-abstract Transformer framework to overcome the problem. Specifically, we leverage pre-trained language models to construct a hierarchical extractor for salient sentence selection across documents and an abstractor for rewriting the selected contents as summaries. However, learning such a framework is challenging since the optimal contents for the abstractor are generally unknown. Previous works typically create *pseudo extraction oracle* to enable the supervised learning for both the extractor and the abstractor. Nevertheless, we argue that the performance of such methods could be restricted due to the insufficient information for prediction and inconsistent objectives between training and testing. To this end, we propose a loss weighting mechanism that makes the model aware of the unequal importance for the sentences not in the pseudo extraction oracle, and leverage the fine-tuned abstractor to generate summary references as auxiliary signals for learning the extractor. Moreover, we propose a reinforcement learning method that can efficiently apply to the extractor for harmonizing the optimization between training and testing. Experiment results show that our framework substantially outperforms strong baselines with comparable model sizes and achieves the best results on the Multi-News, Multi-XScience, and WikiCatSum corpora.

SueNes: A Weakly Supervised Approach to Evaluating Single-Document Summarization via Negative Sampling
Forrest Sheng Bao, Ge Luo, Hebi Li, Minghui Qiu, Yinfei Yang, Youbiao He and Cen Chen 16:15-17:45 (702 Clearwater)

[3] https://github.com/LemonQC/KG2Text

Canonical automatic summary evaluation metrics, such as ROUGE, focus on lexical similarity which cannot well capture semantics nor linguistic quality and require a reference summary which is costly to obtain. Recently, there have been a growing number of efforts to alleviate either or both of the two drawbacks. In this paper, we present a proof-of-concept study to a weakly supervised summary evaluation approach without the presence of reference summaries. Massive data in existing summarization datasets are transformed for training by pairing documents with corrupted reference summaries. In cross-domain tests, our strategy outperforms baselines with promising improvements, and show a great advantage in gauging linguistic qualities over all metrics.

Investigating Crowdsourcing Protocols for Evaluating the Factual Consistency of Summaries
Xiangru Tang, Alexander Fabbri, Haoran Li, Ziming Mao, Griffin Thomas Adams, Borui Wang, Asli Celikyilmaz, Yashar Mehdad and Dragomir Radev 16:15-17:45 (702 Clearwater)
Current pre-trained models applied for summarization are prone to factual inconsistencies that misrepresent the source text. Evaluating the factual consistency of summaries is thus necessary to develop better models. However, the human evaluation setup for evaluating factual consistency has not been standardized. To determine the factors that affect the reliability of the human evaluation, we crowdsource evaluations for factual consistency across state-of-the-art models on two news summarization datasets using the rating-based Likert Scale and ranking-based Best-Worst Scaling. Our analysis reveals that the ranking-based Best-Worst Scaling offers a more reliable measure of summary quality across datasets and that the reliability of Likert ratings highly depends on the target dataset and the evaluation design. To improve crowdsourcing reliability, we extend the scale of the Likert rating and present a scoring algorithm for Best-Worst Scaling that we call value learning. Our crowdsourcing guidelines will be publicly available to facilitate future work on factual consistency in summarization.

Detect Rumors in Microblog Posts for Low-Resource Domains via Adversarial Contrastive Learning
Hongzhan Lin, Jing Ma, Liangliang Chen, Zhiwei Yang, Mingfei Cheng and Chen Guang 16:15-17:45 (702 Clearwater)
Massive false rumors emerging along with breaking news or trending topics severely hinder the truth. Existing rumor detection approaches achieve promising performance on the yesterday's news, since there is enough corpus collected from the same domain for model training. However, they are poor at detecting rumors about unforeseen events especially those propagated in minority languages due to the lack of training data and prior knowledge (i.e., low-resource regimes). In this paper, we propose an adversarial contrastive learning framework to detect rumors by adapting the features learned from well-resourced rumor data to that of the low-resourced. Our model explicitly overcomes the restriction of domain and/or language usage via language alignment and a novel supervised contrastive training paradigm. Moreover, we develop an adversarial augmentation mechanism to further enhance the robustness of low-resource rumor representation. Extensive experiments conducted on two low-resource datasets collected from real-world microblog platforms demonstrate that our framework achieves much better performance than state-of-the-art methods and exhibits a superior capacity for detecting rumors at early stages.

Word Tour: One-dimensional Word Embeddings via the Traveling Salesman Problem
Ryoma Sato 16:15-17:45 (702 Clearwater)
Word embeddings are one of the most fundamental technologies used in natural language processing. Existing word embeddings are high-dimensional and consume considerable computational resources. In this study, we propose WordTour, unsupervised one-dimensional word embeddings. To achieve the challenging goal, we propose a decomposition of the desiderata of word embeddings into two parts, completeness and soundness, and focus on soundness in this paper. Owing to the single dimensionality, WordTour is extremely efficient and provides a minimal means to handle word embeddings. We experimentally confirmed the effectiveness of the proposed method via user study and document classification.

Causal Distillation for Language Models
Zhengxuan Wu, Atticus Geiger, Joshua Rozner, Elisa Kreiss, Hanson Lu, Thomas Icard, Christopher Potts and Noah Goodman 16:15-17:45 (702 Clearwater)
Distillation efforts have led to language models that are more compact and efficient without serious drops in performance. The standard approach to distillation trains a student model against two objectives: a task-specific objective (e.g., language modeling) and an imitation objective that encourages the hidden states of the student model to be similar to those of the larger teacher model. In this paper, we show that it is beneficial to augment distillation with a third objective that encourages the student to imitate the *causal* dynamics of the teacher through a distillation interchange intervention training objective (DIITO). DIITO pushes the student model to become a *causal abstraction* of the teacher model – a faithful model with simpler causal structure. DIITO is fully differentiable, easily implemented, and combines flexibly with other objectives. Compared against standard distillation with the same setting, DIITO results in lower perplexity on the WikiText-103M corpus (masked language modeling) and marked improvements on the GLUE benchmark (natural language understanding), SQuAD (question answering), and CoNLL-2003 (named entity recognition).

Attention Fusion: a light yet efficient late fusion mechanism for task adaptation in NLU
Jin Cao, Chandana Satya Prakash and Wael Hamza 16:15-17:45 (702 Clearwater)
Fine-tuning a pre-trained language model using annotated data has become the de-facto standard for adapting general-purpose pre-trained models like BERT to downstream tasks. However, given the trend of larger pre-trained models, fine-tuning these models for each downstream task is parameter-inefficient and computationally-expensive deeming this approach sub-optimal for adoption by NLU systems. In recent years, various approaches have been proposed for parameter efficient task adaptation such as Adaptor, Bitfit, Prompt tuning, Prefix tuning etc. However, most of these efforts propose to insert task specific parameters in-between or inside intermediate layers of the pre-trained encoder resulting in higher computational cost due to back-propagation of errors to all layers. To mitigate this issue, we propose a light but efficient, attention based fusion module which computes task-attuned token representations by aggregating intermediate layer representations from a pre-trained network. Our proposed fusion module trains only 0.0009% of total parameters and achieves competitive performance to the standard fine-tuning approach on various tasks. It is also decoupled from the pre-trained network making it efficient during computation and scalable during deployment. Last but not the least, we demonstrate that our proposed attention-fusion mechanism can transfer effectively to different languages for further re-use and expansion.

Towards Computationally Feasible Deep Active Learning
Akim Tsvigun, Artem Shelmanov, Gleb Kuzmin, Leonid Sanochkin, Daniil Larionov, Gleb Gennadjevich Gusev, Manvel Avetisian and Leonid Zhukov 16:15-17:45 (702 Clearwater)
Active learning (AL) is a prominent technique for reducing the annotation effort required for training machine learning models. Deep learning offers a solution for several essential obstacles to deploying AL in practice but introduces many others. One of such problems is the excessive computational resources required to train an acquisition model and estimate its uncertainty on instances in the unlabeled pool. We propose two techniques that tackle this issue for text classification and tagging tasks, offering a substantial reduction of AL iteration duration and the computational overhead introduced by deep acquisition models in AL. We also demonstrate that our algorithm that leverages pseudo-labeling and distilled models overcomes one of the essential obstacles revealed previously in the literature. Namely, it was shown that due to differences between an acquisition model used to select instances during AL and a successor model trained on the labeled data, the benefits of AL can diminish. We show that our algorithm, despite using a smaller and faster acquisition model, is capable of training a more expressive successor model with higher performance.

Pruning Adatperfusion with Lottery Ticket Hypothesis
Jiarun Wu, Qingliang Chen, Zeguan Xiao, Yuliang Gu and Mengsi Sun 16:15-17:45 (702 Clearwater)
Pre-trained language models have shown great success in multiple downstream tasks. However, they are computationally expensive to fine-tune. Thus, transfer learning with adapter modules has been introduced to alleviate this problem, helping to extract knowledge of the downstream tasks. Adapterfusion models are an example of the transformers-with-adapter-modules, which merge multiple adapters to incorporate knowledge from different tasks. However, merging multiple adapters will inevitably cause redundancies, increasing the training and inference time massively. Therefore, in this paper, we propose an approach to identify the influence of each adapter module and a novel way to prune adapters based on the prestigious Lottery Ticket Hypothesis. Experiments on GLUE datasets show that the pruned Adapterfusion model with our scheme can achieve state-of-the-art results, reducing sizes significantly while keeping performance intact.

Do Deep Neural Nets Display Human-like Attention in Short Answer Scoring?
Zijie Zeng, XINYU LI, Dragan Gasevic and Guanliang Chen 16:15-17:45 (702 Clearwater)
Deep Learning (DL) techniques have been increasingly adopted for Automatic Text Scoring in education. However, these techniques often suffer from their inabilities to explain and justify how a prediction is made, which, unavoidably, decreases their trustworthiness and hinders educators from embracing them in practice. This study aimed to investigate whether (and to what extent) DL-based graders align with human graders regarding the important words they identify when marking short answer questions. To this end, we first conducted a user study to ask human graders to manually annotate important words in assessing answer quality and then measured the overlap between these human-annotated words and those identified by DL-based graders (i.e., those receiving large attention weights). Furthermore, we ran a randomized controlled experiment to explore the impact of highlighting important words detected by DL-based graders on human grading. The results showed that: (i) DL-based graders, to a certain degree, displayed alignment with human graders no matter whether DL-based graders and human graders agreed on the quality of an answer; and (ii) it is possible to facilitate human grading by highlighting those DL-detected important words, though further investigations are necessary to understand how human graders exploit such highlighted words.

Towards Process-Oriented, Modular, and Versatile Question Generation that Meets Educational Needs
Xu Wang, Simin Fan, Jessica Houghton and Lu Wang 16:15-17:45 (702 Clearwater)
NLP-powered automatic question generation (QG) techniques carry great pedagogical potential of saving educators' time and benefiting student learning. Yet, QG systems have not been widely adopted in classrooms to date. In this work, we aim to pinpoint key impediments and investigate how to improve the usability of automatic QG techniques for educational purposes by understanding how instructors construct questions and identifying touch points to enhance the underlying NLP models. We perform an in-depth need finding study with 11 instructors across 7 different universities, and summarize their thought processes and needs when creating questions. While instructors show great interests in using NLP systems to support question design, none of them has used such tools in practice. They resort to multiple sources of information, ranging from domain knowledge to students' misconceptions, all of which missing from today's QG systems. We argue that building effective human-NLP collaborative QG systems that emphasize instructor control and explainability is imperative for real-world adoption. We call for QG systems to provide process-oriented support, use modular design, and handle diverse sources of input.

Grapheme-to-Phoneme Conversion for Thai using Neural Regression Models
Tomohiro Yamasaki 16:15-17:45 (702 Clearwater)
We propose a novel Thai grapheme-to-phoneme conversion method based on a neural regression model that is trained using neural networks to predict the similarity between a candidate and the correct pronunciation. After generating a set of candidates for an input word or phrase using the orthography rules, this model selects the best-similarity pronunciation from the candidates. This method can be applied to languages other than Thai simply by preparing enough orthography rules, and can reduce the mistakes that neural network models often make. We show that the accuracy of the proposed method is .931, which is comparable to that of encoder-decoder sequence models. We also demonstrate that the proposed method is superior in terms of the difference between correct and predicted pronunciations because incorrect, strange output sometimes occurs when using encoder-decoder sequence models but the error is within the expected range when using the proposed method.

End-to-end Spoken Conversational Question Answering: Task, Dataset and Model
Chenyu You, Nuo Chen, Fenglin Liu, Shen Ge, Xian Wu and Yuexian Zou 16:15-17:45 (702 Clearwater)
In spoken question answering, the systems are designed to answer questions from contiguous text spans within the related speech transcripts. However, the most natural way that human seek or test their knowledge is via human conversations. Therefore, we propose a new Spoken Conversational Question Answering task (SCQA), aiming at enabling the systems to model complex dialogues flow given the speech documents. In this task, our main objective is to build the system to deal with conversational questions based on the audio recordings, and to explore the plausibility of providing more cues from different modalities with systems in information gathering. To this end, instead of directly adopting automatically generated speech transcripts with highly noisy data, we propose a novel unified data distillation approach, DDNet, which effectively ingests cross-modal information to achieve fine-grained representations of the speech and language modalities. Moreover, we propose a simple and novel mechanism, termed Dual Attention, by encouraging better alignments between audio and text to ease the process of knowledge transfer. To evaluate the capacity of SCQA systems in a dialogue-style interaction, we assemble a Spoken Conversational Question Answering (Spoken-CoQA) dataset with more than 40k question-answer pairs from 4k conversations. We first show that the performance of the existing state-of-the-art methods significantly degrade on our dataset, hence demonstrating the necessity of incorporating cross-modal information to achieve good performance gains. Our experimental results demonstrate that our proposed method achieves superior performance in spoken conversational question answering. Codes and datasets will be made publicly available.

Sentence-Level Resampling for Named Entity Recognition
Xiaochen Wang and Yue Wang 16:15-17:45 (702 Clearwater)
As a fundamental task in natural language processing, named entity recognition (NER) aims to locate and classify named entities in unstructured text. However, named entities are always the minority among all tokens in the text. This data imbalance problem presents a challenge to machine learning models as their learning objective is usually dominated by the majority of non-entity tokens. To alleviate data imbalance, we propose a set of sentence-level resampling methods where the importance of each training sentence is computed based on its tokens and entities. We study the generalizability of these resampling methods on a wide variety of NER models (CRF, Bi-LSTM, and BERT) across corpora from diverse domains (general, social, and medical texts). Extensive experiments show that the proposed methods improve span-level macro F1-scores of the evaluated NER models on multiple corpora, frequently outperforming sub-sentence-level resampling, data augmentation, and special loss functions such as focal and Dice loss.

Unified Semantic Typing with Meaningful Label Inference
James Y. Huang, Bangzheng Li, Jiashu Xu and Muhao Chen 16:15-17:45 (702 Clearwater)
Semantic typing aims at classifying tokens or spans of interest in a textual context into semantic categories such as relations, entity types, and event types. The inferred labels of semantic categories meaningfully interpret how machines understand components of text. In this paper, we present UniST, a unified framework for semantic typing that captures label semantics by projecting both inputs and labels into a joint semantic embedding space. To formulate different lexical and relational semantic typing tasks as a unified task, we incorporate task descriptions to be

jointly encoded with the input, allowing UniST to be adapted to different tasks without introducing task-specific model components. UniST optimizes a margin ranking loss such that the semantic relatedness of the input and labels is reflected from their embedding similarity. Our experiments demonstrate that UniST achieves strong performance across three semantic typing tasks: entity typing, relation classification and event typing. Meanwhile, UniST effectively transfers semantic knowledge of labels and substantially improves generalizability on inferring rarely seen and unseen types. In addition, multiple semantic typing tasks can be jointly trained within the unified framework, leading to a single compact multi-tasking model that performs comparably to dedicated single-task models, while offering even better transferability.

Crossroads, Buildings and Neighborhoods: A Dataset for Fine-grained Location Recognition
Pei Chen, Haotian Xu, Cheng Zhang and Ruihong Huang 16:15-17:45 (702 Clearwater)
General domain Named Entity Recognition (NER) datasets like CoNLL-2003 mostly annotate coarse-grained location entities such as a country or a city. But many applications require identifying fine-grained locations from texts and mapping them precisely to geographic sites, e.g., a crossroad, an apartment building, or a grocery store. In this paper, we introduce a new dataset HarveyNER with fine-grained locations annotated in tweets. This dataset presents unique challenges and characterizes many complex and long location mentions in informal descriptions. We built strong baseline models using Curriculum Learning and experimented with different heuristic curricula to better recognize difficult location mentions. Experimental results show that the simple curricula can improve the system's performance on hard cases and its overall performance, and outperform several other baseline systems. The dataset and the baseline models can be found at https://github.com/brickee/HarveyNER.

Modeling Task Interactions in Document-Level Joint Entity and Relation Extraction
Liyan Xu and Jinho D. Choi 16:15-17:45 (702 Clearwater)
We target on the document-level relation extraction in an end-to-end setting, where the model needs to jointly perform mention extraction, coreference resolution (COREF) and relation extraction (RE) at once, and gets evaluated in an entity-centric way. Especially, we address the two-way interaction between COREF and RE that has not been the focus by previous work, and propose to introduce explicit interaction namely Graph Compatibility (GC) that is specifically designed to leverage task characteristics, bridging decisions of two tasks for direct task interference. Our experiments are conducted on DocRED and DWIE; in addition to GC, we implement and compare different multi-task settings commonly adopted in previous work, including pipeline, shared encoders, graph propagation, to examine the effectiveness of different interactions. The result shows that GC achieves the best performance by up to 2.3/5.1 F1 improvement over the baseline.

Towards Process-Oriented, Modular, and Versatile Question Generation that Meets Educational Needs
Xu Wang, Simin Fan, Jessica Houghton and Lu Wang 16:15-17:45 (702 Clearwater)
NLP-powered automatic question generation (QG) techniques carry great pedagogical potential of saving educators' time and benefiting student learning. Yet, QG systems have not been widely adopted in classrooms to date. In this work, we aim to pinpoint key impediments and investigate how to improve the usability of automatic QG techniques for educational purposes by understanding how instructors construct questions and identifying touch points to enhance the underlying NLP models. We perform an in-depth need finding study with 11 instructors across 7 different universities, and summarize their thought processes and needs when creating questions. While instructors show great interests in using NLP systems to support question design, none of them has used such tools in practice. They resort to multiple sources of information, ranging from domain knowledge to students' misconceptions, all of which missing from today's QG systems. We argue that building effective human-NLP collaborative QG systems that emphasize instructor control and explainability is imperative for real-world adoption. We call for QG systems to provide process-oriented support, use modular design, and handle diverse sources of input.

Improving Neural Models for Radiology Report Retrieval with Lexicon-based Automated Annotation
Luyao Shi, Tanveer Syeda-mahmood and Tyler Baldwin 16:15-17:45 (702 Clearwater)
Many clinical informatics tasks that are based on electronic health records (EHR) need relevant patient cohorts to be selected based on findings, symptoms and diseases. Frequently, these conditions are described in radiology reports which can be retrieved using information retrieval (IR) methods. The latest of these techniques utilize neural IR models such as BERT trained on clinical text. However, these methods still lack semantic understanding of the underlying clinical conditions as well as ruled out findings, resulting in poor precision during retrieval. In this paper we combine clinical finding detection with supervised query match learning. Specifically, we use lexicon-driven concept detection to detect relevant findings in sentences. These findings are used as queries to train a Sentence-BERT (SBERT) model using triplet loss on matched and unmatched query-sentence pairs. We show that the proposed supervised training task remarkably improves the retrieval performance of SBERT. The trained model generalizes well to unseen queries and reports from different collections.

Is Neural Topic Modelling Better than Clustering? An Empirical Study on Clustering with Contextual Embeddings for Topics
Zihan Zhang, Meng Fang, Ling Chen and Mohammad Reza Namazi Rad 16:15-17:45 (702 Clearwater)
Recent work incorporates pre-trained word embeddings such as BERT embeddings into Neural Topic Models (NTMs), generating highly coherent topics. However, with high-quality contextualized document representations, do we really need sophisticated neural models to obtain coherent and interpretable topics? In this paper, we conduct thorough experiments showing that directly clustering high-quality sentence embeddings with an appropriate word selecting method can generate more coherent and diverse topics than NTMs, achieving also higher efficiency and simplicity.

Cross-document Misinformation Detection based on Event Graph Reasoning
Xueqing Wu, Kung-Hsiang Huang, Yi Fung and Heng Ji 16:15-17:45 (702 Clearwater)
For emerging events, human readers are often exposed to both real news and fake news. Multiple news articles may contain complementary or contradictory information that readers can leverage to help detect fake news. Inspired by this process, we propose a novel task of cross-document misinformation detection. Given a cluster of topically related news documents, we aim to detect misinformation at both document level and a more fine-grained level, event level. Due to the lack of data, we generate fake news by manipulating real news, and construct 3 new datasets with 422, 276, and 1,413 clusters of topically related documents, respectively. We further propose a graph-based detector that constructs a cross-document knowledge graph using cross-document event coreference resolution and employs a heterogeneous graph neural network to conduct detection at two levels. We then feed the event-level detection results into the document-level detector. Experimental results show that our proposed method significantly outperforms existing methods by up to 7 F1 points on this new task.

A Word is Worth A Thousand Dollars: Adversarial Attack on Tweets Fools Stock Prediction
Yong Xie, Dakuo Wang, Pin-Yu Chen, Jinjun Xiong, Sijia Liu and Oluwasanmi O Koyejo 16:15-17:45 (702 Clearwater)
More and more investors and machine learning models rely on social media (e.g., Twitter and Reddit) to gather information and predict movements stock prices. Although text-based models are known to be vulnerable to adversarial attacks, whether stock prediction models have similar vulnerability given necessary constraints is underexplored. In this paper, we experiment with a variety of adversarial attack configurations to fool three stock prediction victim models. We address the task of adversarial generation by solving combinatorial optimization problems with semantics and budget constraints. Our results show that the proposed attack method can achieve consistent success rates and cause significant monetary loss in trading simulation by simply concatenating a perturbed but semantically similar tweet.

Privacy-Preserving Text Classification on BERT Embeddings with Homomorphic Encryption
Garam Lee, Minsoo Kim, Jai Hyun Park, Seung-won Hwang and Jung Hee Cheon 16:15-17:45 (702 Clearwater)
Embeddings, which compress information in raw text into semantics-preserving low-dimensional vectors, have been widely adopted for their efficacy. However, recent research has shown that embeddings can potentially leak private information about sensitive attributes of the text, and in some cases, can be inverted to recover the original input text. To address these growing privacy challenges, we propose a privatization mechanism for embeddings based on homomorphic encryption, to prevent potential leakage of any piece of information in the process of text classification. In particular, our method performs text classification on the encryption of embeddings from state-of-the-art models like BERT, supported by an efficient GPU implementation of CKKS encryption scheme. We show that our method offers encrypted protection of BERT embeddings, while largely preserving their utility on downstream text classification tasks.

Harmless Transfer Learning for Item Embeddings
Chengyue Gong, Xiaocong Du, Dhruv Choudhary, Bhargav Bhushanam, Qiang Liu and Arun Kejariwal 16:15-17:45 (702 Clearwater)
Learning embedding layers (for classes, words, items, etc.) is a key component of lots of applications, ranging from natural language processing, recommendation systems to electronic health records, etc. However, the frequency of real-world items follows a long-tail distribution in these applications, causing naive training methods perform poorly on the rare items. A line of previous works address this problem by transferring the knowledge from the frequent items to rare items by introducing an auxiliary transfer loss. However, when defined improperly, the transfer loss may introduce harmful biases and deteriorate the performance.

In this work, we propose a harmless transfer learning framework that limits the impact of the potential biases in both the definition and optimization of the transfer loss. On the definition side, we reduce the bias in transfer loss by focusing on the items to which information from high-frequency items can be efficiently transferred. On the optimization side, we leverage a lexicographic optimization framework to efficiently incorporate the information of the transfer loss without hurting the minimization of the main prediction loss function. Our method serves as a plug-in module and significantly boosts the performance on a variety of NLP and recommendation system tasks.

A Robustly Optimized BMRC for Aspect Sentiment Triplet Extraction
Shu Liu, Kaiwen Li and Zuhe Li 16:15-17:45 (702 Clearwater)
Aspect sentiment triplet extraction (ASTE) is a challenging subtask in aspect-based sentiment analysis. It aims to explore the triplets of aspects, opinions and sentiments with complex correspondence from the context. The bidirectional machine reading comprehension (BMRC), can effectively deal with ASTE task, but several problems remains, such as query conflict and probability unilateral decrease. Therefore, this paper presents a robustly optimized BMRC method by incorporating four improvements. The word segmentation is applied to facilitate the semantic learning. Exclusive classifiers are designed to avoid the interference between different queries. A span matching rule is proposed to select the aspects and opinions that better represent the expectations of the model. The probability generation strategy is also introduced to obtain the predicted probability for aspects, opinions and aspect-opinion pairs. We have conducted extensive experiments on multiple benchmark datasets, where our model achieves the state-of-the-art performance.

Data Augmentation with Dual Training for Offensive Span Detection
Nasim Nouri 16:15-17:45 (702 Clearwater)
Recognizing offensive text is an important requirement for every content management system, especially for social networks. While the majority of the prior work formulate this problem as text classification, i.e., if a text excerpt is offensive or not, in this work we propose a novel model for offensive span detection (OSD), whose goal is to identify the spans responsible for the offensive tone of the text. One of the challenges to train a model for this novel setting is the lack of enough training data. To address this limitation, in this work we propose a novel method in which the large-scale pre-trained language model GPT-2 is employed to generate synthetic training data for OSD. In particular, we propose to train the GPT-2 model in a dual-training setting using the REINFORCE algorithm to generate in-domain, natural and diverse training samples. Extensive experiments on the benchmark dataset for OSD reveal the effectiveness of the proposed method.

Multi-Domain Targeted Sentiment Analysis
Orith Toledo-Ronen, Matan Orbach, Yoav Katz and Noam Slonim 16:15-17:45 (702 Clearwater)
Targeted Sentiment Analysis (TSA) is a central task for generating insights from consumer reviews. Such content is extremely diverse, with sites like Amazon or Yelp containing reviews on products and businesses from many different domains. A real-world TSA system should gracefully handle that diversity. This can be achieved by a multi-domain model – one that is robust to the domain of the analyzed texts, and performs well on various domains. To address this scenario, we present a multi-domain TSA system based on augmenting a given training set with diverse weak labels from assorted domains. These are obtained through self-training on the Yelp reviews corpus. Extensive experiments with our approach on three evaluation datasets across different domains demonstrate the effectiveness of our solution. We further analyze how restrictions imposed on the available labeled data affect the performance, and compare the proposed method to the costly alternative of manually gathering diverse TSA labeled data. Our results and analysis show that our approach is a promising step towards a practical domain-robust TSA system.

UserIdentifier: Implicit User Representations for Simple and Effective Personalized Sentiment Analysis
*Fatemehsadat Mireshghallah, Vaishnavi Shrivastava, Milad Shokouhi, Taylor Berg-Kirkpatrick, Robert Sim and Dimitrios Dimitriadis*16:15-17:45 (702 Clearwater)
Global models are typically trained to be as generalizable as possible. Invariance to the specific user is considered desirable since models are shared across multitudes of users. However, these models are often unable to produce personalized responses for individual users, based on their data. Contrary to widely-used personalization techniques based on few-shot and meta-learning, we propose UserIdentifier, a novel scheme for training a single shared model for all users. Our approach produces personalized responses by prepending a fixed, user-specific non-trainable string (called “user identifier”) to each user’s input text. Unlike prior work, this method doesn’t need any additional model parameters, any extra rounds of personal few-shot learning or any change made to the vocabulary. We empirically study different types of user identifiers (numeric, alphanumeric, and also randomly generated) and demonstrate that, surprisingly, randomly generated user identifiers outperform the prefix-tuning based state-of-the-art approach by up to 13, on a suite of sentiment analysis datasets.

Semantic Diversity in Dialogue with Natural Language Inference
Katherine Stasaski and Marti Hearst 16:15-17:45 (702 Clearwater)
Generating diverse, interesting responses to chitchat conversations is a problem for neural conversational agents. This paper makes two substantial contributions to improving diversity in dialogue generation. First, we propose a novel metric which uses Natural Language Inference (NLI) to measure the semantic diversity of a set of model responses for a conversation. We evaluate this metric using an established framework (Tevet and Berant, 2021) and find strong evidence indicating NLI Diversity is correlated with semantic diversity. Specifically, we show that the contradiction relation is more useful than the neutral relation for measuring this diversity and that incorporating the NLI model’s confidence achieves state-of-the-art results. Second, we demonstrate how to iteratively improve the semantic diversity of a sampled set of responses via a new generation procedure called Diversity Threshold Generation, which results in an average 137% increase in NLI Diversity compared to standard generation procedures.

CS1QA: A Dataset for Assisting Code-based Question Answering in an Introductory Programming Course
Changyoon Lee, Yeon Seonwoo and Alice Oh 16:15-17:45 (702 Clearwater)
We introduce CS1QA, a dataset for code-based question answering in the programming education domain. CS1QA consists of 9,237 question-answer pairs gathered from chat logs in an introductory programming class using Python, and 17,698 unannotated chat data with code. Each question is accompanied with the student's code, and the portion of the code relevant to answering the question. We carefully design the annotation process to construct CS1QA, and analyze the collected dataset in detail. The tasks for CS1QA are to predict the question type, the relevant code snippet given the question and the code and retrieving an answer from the annotated corpus. Results for the experiments on several baseline models are reported and thoroughly analyzed. The tasks for CS1QA challenge models to understand both the code and natural language. This unique dataset can be used as a benchmark for source code comprehension and question answering in the educational setting.

The USMLE® Step 2 Clinical Skills Patient Note Corpus
Victoria Yaneva, Janet Mee, Le An Ha, Polina Harik, Michael Jodoin and Alex J Mechaber 16:15-17:45 (702 Clearwater)
This paper presents a corpus of 43,985 clinical patient notes (PNs) written by 35,156 examinees during the high-stakes USMLE® Step 2 Clinical Skills examination. In this exam, examinees interact with standardized patients - people trained to portray simulated scenarios called clinical cases. For each encounter, an examinee writes a PN, which is then scored by physician raters using a rubric of clinical concepts, expressions of which should be present in the PN. The corpus features PNs from 10 clinical cases, as well as the clinical concepts from the case rubrics. A subset of 2,840 PNs were annotated by 10 physician experts such that all 143 concepts from the case rubrics (e.g., shortness of breath) were mapped to 34,660 PN phrases (e.g., dyspnea, difficulty breathing). The corpus is available via a data sharing agreement with NBME and can be requested at https://www.nbme.org/services/data-sharing.

Transparent Human Evaluation for Image Captioning
Jungo Kasai, Keisuke Sakaguchi, Lavinia Dunagan, Jacob Daniel Morrison, Ronan Le Bras, Yejin Choi and Noah Smith 16:15-17:45 (702 Clearwater)
We establish THumB, a rubric-based human evaluation protocol for image captioning models. Our scoring rubrics and their definitions are carefully developed based on machine- and human-generated captions on the MSCOCO dataset. Each caption is evaluated along two main dimensions in a tradeoff (precision and recall) as well as other aspects that measure the text quality (fluency, conciseness, and inclusive language). Our evaluations demonstrate several critical problems of the current evaluation practice. Human-generated captions show substantially higher quality than machine-generated ones, especially in coverage of salient information (i.e., recall), while most automatic metrics say the opposite. Our rubric-based results reveal that CLIPScore, a recent metric that uses image features, better correlates with human judgments than conventional text-only metrics because it is more sensitive to recall. We hope that this work will promote a more transparent evaluation protocol for image captioning and its automatic metrics.

ChapterBreak: A Challenge Dataset for Long-Range Language Models
Simeng Sun, Katherine Thai and Mohit Iyyer 16:15-17:45 (702 Clearwater)
While numerous architectures for long-range language models (LRLMs) have recently been proposed, a meaningful evaluation of their discourse-level language understanding capabilities has not yet followed. To this end, we introduce ChapterBreak, a challenge dataset that provides an LRLM with a long segment from a narrative that ends at a chapter boundary and asks it to distinguish the beginning of the ground-truth next chapter from a set of negative segments sampled from the same narrative. A fine-grained human annotation reveals that our dataset contains many complex types of chapter transitions (e.g., parallel narratives, cliffhanger endings) that require processing global context to comprehend. Experiments on ChapterBreak show that existing LRLMs fail to effectively leverage long-range context, substantially underperforming a segment-level model trained directly for this task. We publicly release our ChapterBreak dataset to spur more principled future research into LRLMs.

TVShowGuess: Character Comprehension in Stories as Speaker Guessing
Yisi Sang, Xiangyang Mou, Mo Yu, Shunyu Yao, Jing Li and Jeffrey Stanton 16:15-17:45 (702 Clearwater)
We propose a new task for assessing machines' skills of understanding fictional characters in narrative stories. The task, TVShowGuess, builds on the scripts of TV series and takes the form of guessing the anonymous main characters based on the backgrounds of the scenes and the dialogues. Our human study supports that this form of task covers comprehension of multiple types of character persona, including understanding characters' personalities, facts and memories of personal experience, which are well aligned with the psychological and literary theories about the theory of mind (ToM) of human beings on understanding fictional characters during reading. We further propose new model architectures to support the contextualized encoding of long scene texts. Experiments show that our proposed approaches significantly outperform baselines, yet still largely lag behind the (nearly perfect) human performance. Our work serves as a first step toward the goal of narrative character comprehension.

Building Multilingual Machine Translation Systems That Serve Arbitrary XY Translations
Akiko Eriguchi, Shufang Xie, Tao Qin and Hany Hassan 16:15-17:45 (702 Clearwater)
Multilingual Neural Machine Translation (MNMT) enables one system to translate sentences from multiple source languages to multiple target languages, greatly reducing deployment costs compared with conventional bilingual systems. The MNMT training benefit, however, is often limited to many-to-one directions. The model suffers from poor performance in one-to-many and many-to-many with zero-shot setup. To address this issue, this paper discusses how to practically build MNMT systems that serve arbitrary X-Y translation directions while leveraging multilinguality with a two-stage training strategy of pretraining and finetuning. Experimenting with the WMT'21 multilingual translation task, we demonstrate that our systems outperform the conventional baselines of direct bilingual models and pivot translation models for most directions, averagely giving +6.0 and +4.1 BLEU, without the need for architecture change or extra data collection. Moreover, we also examine our proposed approach in an extremely large-scale data setting to accommodate practical deployment scenarios.

Quality-Aware Decoding for Neural Machine Translation
Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. De Souza, Perez Ogayo, Graham Neubig and Andre Martins 16:15-17:45 (702 Clearwater)
Despite the progress in machine translation quality estimation and evaluation in the last years, decoding in neural machine translation (NMT) is mostly oblivious to this and centers around finding the most probable translation according to the model (MAP decoding), approximated with beam search. In this paper, we bring together these two lines of research and propose *quality-aware decoding* for NMT, by leveraging recent breakthroughs in reference-free and reference-based MT evaluation through various inference methods like N-best reranking and minimum Bayes risk decoding. We perform an extensive comparison of various possible candidate generation and ranking methods across four datasets and two model classes and find that quality-aware decoding consistently outperforms MAP-based decoding according both to state-of-the-art automatic metrics (COMET and BLEURT) and to human assessments.

A Study of Syntactic Multi-Modality in Non-Autoregressive Machine Translation
Kexun Zhang, Rui Wang, Xu Tan, Junliang Guo, Yi Ren, Tao Qin and Tie-Yan Liu 16:15-17:45 (702 Clearwater)

It is difficult for non-autoregressive translation (NAT) models to capture the multi-modal distribution of target translations due to their conditional independence assumption, which is known as the "multi-modality problem", including the lexical multi-modality and the syntactic multi-modality. While the first one has been well studied, the syntactic multi-modality brings severe challenges to the standard cross entropy (XE) loss in NAT and is understudied. In this paper, we conduct a systematic study on the syntactic multi-modality problem. Specifically, we decompose it into short- and long-range syntactic multi-modalities and evaluate several recent NAT algorithms with advanced loss functions on both carefully designed synthesized datasets and real datasets. We find that the Connectionist Temporal Classification (CTC) loss and the Order-Agnostic Cross Entropy (OAXE) loss can better handle short- and long-range syntactic multi-modalities respectively. Furthermore, we take the best of both and design a new loss function to better handle the complicated syntactic multi-modality in real-world datasets. To facilitate practical usage, we provide a guide to using different loss functions for different kinds of syntactic multi-modality.

Tricks for Training Sparse Translation Models
Dheeru Dua, Shruti Bhosale, Vedanuj Goswami, James Cross, Mike Lewis and Angela Fan 16:15-17:45 (702 Clearwater)
Multi-task learning with an unbalanced data distribution skews model learning towards high resource tasks, especially when model capacity is fixed and fully shared across all tasks. Sparse scaling architectures, such as BASELayers, provide flexible mechanisms for different tasks to have a variable number of parameters, which can be useful to counterbalance skewed data distributions. We find that that sparse architectures for multilingual machine translation can perform poorly out of the box and propose two straightforward techniques to mitigate this — a temperature heating mechanism and dense pre-training. Overall, these methods improve performance on two multilingual translation benchmarks compared to standard BASELayers and Dense scaling baselines, and in combination, more than 2x model convergence speed.

When do Contrastive Word Alignments Improve Many-to-many Neural Machine Translation?
Zhuoyuan Mao, Chenhui Chu, Raj Dabre, Haiyue Song, Zhen Wan and Sadao Kurohashi 16:15-17:45 (702 Clearwater)
Word alignment has proven to benefit many-to-many neural machine translation (NMT). However, high-quality ground-truth bilingual dictionaries were used for pre-editing in previous methods, which are unavailable for most language pairs. Meanwhile, the contrastive objective can implicitly utilize automatically learned word alignment, which has not been explored in many-to-many NMT. This work proposes a word-level contrastive objective to leverage word alignments for many-to-many NMT. Empirical results show that this leads to 0.8 BLEU gains for several language pairs. Analyses reveal that in many-to-many NMT, the encoder's sentence retrieval performance highly correlates with the translation quality, which explains when the proposed method impacts translation. This motivates future exploration for many-to-many NMT to improve the encoder's sentence retrieval performance.

Reframing Human-AI Collaboration for Generating Free-Text Explanations
Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl and Yejin Choi 16:15-17:45 (702 Clearwater)
Large language models are increasingly capable of generating fluent-appearing text with relatively little task-specific supervision. But can these models accurately explain classification decisions? We consider the task of generating free-text explanations using human-written examples in a few-shot manner. We find that (1) authoring higher quality prompts results in higher quality generations; and (2) surprisingly, in a head-to-head comparison, crowdworkers often prefer explanations generated by GPT-3 to crowdsourced explanations in existing datasets. Our human studies also show, however, that while models often produce factual, grammatical, and sufficient explanations, they have room to improve along axes such as providing novel information and supporting the label. We create a pipeline that combines GPT-3 with a supervised filter that incorporates binary acceptability judgments from humans in the loop. Despite the intrinsic subjectivity of acceptability judgments, we demonstrate that acceptability is partially correlated with various fine-grained attributes of explanations. Our approach is able to consistently filter GPT-3-generated explanations deemed acceptable by humans.

Implicit n-grams Induced by Recurrence
Xiaobing Sun and Wei Lu 16:15-17:45 (702 Clearwater)
Although self-attention based models such as Transformers have achieved remarkable successes on natural language processing (NLP) tasks, recent studies reveal that they have limitations on modeling sequential transformations (Hahn, 2020), which may prompt re-examinations of recurrent neural networks (RNNs) that demonstrated impressive results on handling sequential data. Despite many prior attempts to interpret RNNs, their internal mechanisms have not been fully understood, and the question on how exactly they capture sequential features remains largely unclear. In this work, we present a study that shows there actually exist some explainable components that reside within the hidden states, which are reminiscent of the classical n-grams features. We evaluated such extracted explainable features from trained RNNs on downstream sentiment analysis tasks and found they could be used to model interesting linguistic phenomena such as negation and intensification. Furthermore, we examined the efficacy of using such n-gram components alone as encoders on tasks such as sentiment analysis and language modeling, revealing they could be playing important roles in contributing to the overall performance of RNNs. We hope our findings could add interpretability to RNN architectures, and also provide inspirations for proposing new architectures for sequential data.

Locally Aggregated Feature Attribution on Natural Language Model Understanding
Sheng Zhang, Jin Wang, Haitao Jiang and Rui Song 16:15-17:45 (702 Clearwater)
With the growing popularity of deep-learning models, model understanding becomes more important. Much effort has been devoted to demystify deep neural networks for better explainability. Some feature attribution methods have shown promising results in computer vision, especially the gradient-based methods where effectively smoothing the gradients with reference data is the key to a robust and faithful result. However, direct application of these gradient-based methods to NLP tasks is not trivial due to the fact that the input consists of discrete tokens and the "reference" tokens are not explicitly defined. In this work, we propose Locally Aggregated Feature Attribution (LAFA), a novel gradient-based feature attribution method for NLP models. Instead of relying on obscure reference tokens, it smooths gradients by aggregating similar reference texts derived from language model embeddings. For evaluation purpose, we also design experiments on different NLP tasks including Entity Recognition and Sentiment Analysis on public datasets and key words detection on constructed Amazon catalogue dataset. The superior performance of the proposed method is demonstrated through experiments.

White-box Testing of NLP models with Mask Neuron Coverage
Arshdeep Sekhon, Yangfeng Ji, Matthew Dwyer and Yanjun Qi 16:15-17:45 (702 Clearwater)
Recent literature has seen growing interest in using black-box strategies like CheckList for testing the behavior of NLP models. Research on white-box testing has developed a number of methods for evaluating how thoroughly the internal behavior of deep models is tested, but they are not applicable to NLP models. We propose a set of white-box testing methods that are customized for transformer-based NLP models. These include MASK NEURON COVERAGE (MNCOVER) that measures how thoroughly the attention layers in models are exercised during testing. We show that MNCOVER can refine testing suites generated by CheckList by substantially reduce them in size, for more than 60% on average, while retaining failing tests – thereby concentrating the fault detection power of the test suite. Further we show how method can be used to guide CheckList input generation, evaluate alternative NLP testing methods, and drive data augmentation to improve accuracy.

Improving Contextual Representation with Gloss Regularized Pre-training
Yu Lin, Zhecheng An, Peihao Wu and Zejun MA 16:15-17:45 (702 Clearwater)
Though achieving impressive results on many NLP tasks, the BERT-like masked language models (MLM) encounter the discrepancy between

pre-training and inference. In light of this gap, we investigate the contextual representation of pre-training and inference from the perspective of word probability distribution. We discover that BERT risks neglecting the contextual word similarity in pre-training. To tackle this issue, we propose an auxiliary gloss regularizer module to BERT pre-training (GR-BERT), to enhance word semantic similarity. By predicting masked words and aligning contextual embeddings to corresponding glosses simultaneously, the word similarity can be explicitly modeled. We design two architectures for GR-BERT and evaluate our model in downstream tasks. Experimental results show that the gloss regularizer benefits BERT in word-level and sentence-level semantic representation. The GR-BERT achieves new state-of-the-art in lexical substitution task and greatly promotes BERT sentence representation in both unsupervised and supervised STS tasks.

SUBS: Subtree Substitution for Compositional Semantic Parsing
Jingfeng Yang, Le Zhang and Diyi Yang 16:15-17:45 (702 Clearwater)
Although sequence-to-sequence models often achieve good performance in semantic parsing for i.i.d. data, their performance is still inferior in compositional generalization. Several data augmentation methods have been proposed to alleviate this problem. However, prior work only leveraged superficial grammar or rules for data augmentation, which resulted in limited improvement. We propose to use subtree substitution for compositional data augmentation, where we consider subtrees with similar semantic functions as exchangeable. Our experiments showed that such augmented data led to significantly better performance on Scan and GeoQuery, and reached new SOTA on compositional split of GeoQuery.

CoSe-Co: Text Conditioned Generative CommonSense Contextualizer
Rachit Bansal, Milan Aggarwal, Sumit Bhatia, Jivat Neet Kaur and Balaji Krishnamurthy 16:15-17:45 (702 Clearwater)
Pre-trained Language Models (PTLMs) have been shown to perform well on natural language tasks. Many prior works have leveraged structured commonsense present in the form of entities linked through labeled relations in Knowledge Graphs (KGs) to assist PTLMs. Retrieval approaches use KG as a separate static module which limits coverage since KGs contain finite knowledge. Generative methods train PTLMs on KG triples to improve the scale at which knowledge can be obtained. However, training on symbolic KG entities limits their applicability in tasks involving natural language text where they ignore overall context. To mitigate this, we propose a CommonSense Contextualizer (CoSe-Co) conditioned on sentences as input to make it generically usable in tasks for generating knowledge relevant to the overall context of input text. To train CoSe-Co, we propose a novel dataset comprising of sentence and commonsense knowledge pairs. The knowledge inferred by CoSe-Co is diverse and contain novel entities not present in the underlying KG. We augment generated knowledge in Multi-Choice QA and Open-ended CommonSense Reasoning tasks leading to improvements over current best methods on CSQA, ARC, QASC and OBQA datasets. We also demonstrate its applicability in improving performance of a baseline model for paraphrase generation task.

MuCPAD: A Multi-Domain Chinese Predicate-Argument Dataset
Yahui Liu, Haoping Yang, Chen Gong, Qingrong Xia, Zhenghua Li and Min Zhang 16:15-17:45 (702 Clearwater)
During the past decade, neural network models have made tremendous progress on in-domain semantic role labeling (SRL). However, performance drops dramatically under the out-of-domain setting. In order to facilitate research on cross-domain SRL, this paper presents MuCPAD, a multi-domain Chinese predicate-argument dataset, which consists of 30,897 sentences and 92,051 predicates from six different domains. MuCPAD exhibits three important features. 1) Based on a frame-free annotation methodology, we avoid writing complex frames for new predicates. 2) We explicitly annotate omitted core arguments to recover more complete semantic structure, considering that omission of content words is ubiquitous in multi-domain Chinese texts. 3) We compile 53 pages of annotation guidelines and adopt strict double annotation for improving data quality. This paper describes in detail the annotation methodology and annotation process of MuCPAD, and presents in-depth data analysis. We also give benchmark results on cross-domain SRL based on MuCPAD.

MGIMN: Multi-Grained Interactive Matching Network for Few-shot Text Classification
Jianhai Zhang, Mieradilijiang Maimaiti, Gao Xing, Yuanhang Zheng and Ji Zhang 16:15-17:45 (702 Clearwater)
Text classification struggles to generalize to unseen classes with very few labeled text instances per class. In such a few-shot learning (FSL) setting, metric-based meta-learning approaches have shown promising results. Previous studies mainly aim to derive a prototype representation for each class. However, they neglect that it is challenging-yet-unnecessary to construct a compact representation which expresses the entire meaning for each class. They also ignore the importance to capture the inter-dependency between query and the support set for few-shot text classification. To deal with these issues, we propose a meta-learning based method MGIMN which performs instance-wise comparison followed by aggregation to generate class-wise matching vectors instead of prototype learning. The key of instance-wise comparison is the interactive matching within the class-specific context and episode-specific context. Extensive experiments demonstrate that the proposed method significantly outperforms the existing SOTA approaches, under both the standard FSL and generalized FSL settings.

DocAMR: Multi-Sentence AMR Representation and Evaluation
Tahira Naseem, Austin Blodgett, Sadhana Kumaravel, Tim O'Gorman, Young-Suk Lee, Jeffrey Flanigan, Ramón Fernandez Astudillo, Radu Florian, Salim Roukos and Nathan Schneider 16:15-17:45 (702 Clearwater)
Despite extensive research on parsing of English sentences into Abstract Meaning Representation (AMR) graphs, which are compared to gold graphs via the Smatch metric, full-document parsing into a unified graph representation lacks well-defined representation and evaluation. Taking advantage of a super-sentential level of coreference annotation from previous work, we introduce a simple algorithm for deriving a unified graph representation, avoiding the pitfalls of information loss from over-merging and lack of coherence from under merging. Next, we describe improvements to the Smatch metric to make it tractable for comparing document-level graphs and use it to re-evaluate the best published document-level AMR parser. We also present a pipeline approach combining the top-performing AMR parser and coreference resolution systems, providing a strong baseline for future research.

Improving negation detection with negation-focused pre-training
Thinh Hung Truong, Timothy Baldwin, Trevor Cohn and Karin Verspoor 16:15-17:45 (702 Clearwater)
Negation is a common linguistic feature that is crucial in many language understanding tasks, yet it remains a hard problem due to diversity in its expression in different types of text. Recent works show that state-of-the-art NLP models underperform on samples containing negation in various tasks, and that negation detection models do not transfer well across domains. We propose a new negation-focused pre-training strategy, involving targeted data augmentation and negation masking, to better incorporate negation information into language models. Extensive experiments on common benchmarks show that our proposed approach improves negation detection performance and generalizability over the strong baseline NegBERT (Khandelwal and Sawant, 2020).

Does Pre-training Induce Systematic Inference? How Masked Language Models Acquire Commonsense Knowledge
Ian Porada, Alessandro Sordoni and Jackie CK Cheung 16:15-17:45 (702 Clearwater)
Transformer models pre-trained with a masked-language-modeling objective (e.g., BERT) encode commonsense knowledge as evidenced by behavioral probes; however, the extent to which this knowledge is acquired by systematic inference over the semantics of the pre-training corpora is an open question. To answer this question, we selectively inject verbalized knowledge into the pre-training minibatches of BERT and evaluate how well the model generalizes to supported inferences after pre-training on the injected knowledge. We find generalization does not improve over the course of pre-training BERT from scratch, suggesting that commonsense knowledge is acquired from surface-level,

co-occurrence patterns rather than induced, systematic reasoning.

Partial-input baselines show that NLI models can ignore context, but they don't.
Neha Srikanth and Rachel Rudinger 16:15-17:45 (702 Clearwater)
When strong partial-input baselines reveal artifacts in crowdsourced NLI datasets, the performance of full-input models trained on such datasets is often dismissed as reliance on spurious correlations. We investigate whether state-of-the-art NLI models are capable of overriding default inferences made by a partial-input baseline. We introduce an evaluation set of 600 examples consisting of perturbed premises to examine a RoBERTa model's sensitivity to edited contexts. Our results indicate that NLI models are still capable of learning to condition on context—a necessary component of inferential reasoning—despite being trained on artifact-ridden datasets.

Analytical Reasoning of Text
Wanjun Zhong, Siyuan Wang, Duyu Tang, Zenan Xu, Daya Guo, Yining Chen, Jiahai Wang, Jian Yin, Ming Zhou and Nan Duan 16:15-17:45 (702 Clearwater)
Analytical reasoning is an essential and challenging task that requires a system to analyze a scenario involving a set of particular circumstances and perform reasoning over it to make conclusions. However, current neural models with implicit reasoning ability struggle to solve this task. In this paper, we study the challenge of analytical reasoning of text and collect a new dataset consisting of questions from the Law School Admission Test from 1991 to 2016. We analyze what knowledge understanding and reasoning abilities are required to do well on this task, and present an approach dubbed ARM. It extracts knowledge such as participants and facts from the context. Such knowledge are applied to an inference engine to deduce legitimate solutions for drawing conclusions. In our experiments, we find that ubiquitous pre-trained models struggle to deal with this task as their performance is close to random guess. Results show that ARM outperforms pre-trained models significantly. Moreover, we demonstrate that ARM has better explicit interpretable reasoning ability.

ATP: AMRize Then Parse! Enhancing AMR Parsing with PseudoAMRs
Liang Chen, Peiyi Wang, Runxin Xu, Tianyu Liu, Zhifang Sui and Baobao Chang 16:15-17:45 (702 Clearwater)
As Abstract Meaning Representation (AMR) implicitly involves compound semantic annotations, we hypothesize auxiliary tasks which are semantically or formally related can better enhance AMR parsing. We find that 1) Semantic role labeling (SRL) and dependency parsing (DP), would bring more performance gain than other tasks e.g. MT and summarization in the text-to-AMR transition even with much less data. 2) To make a better fit for AMR, data from auxiliary tasks should be properly "AMRized" to PseudoAMR before training. Knowledge from shallow level parsing tasks can be better transferred to AMR Parsing with structure transform. 3) Intermediate-task learning is a better paradigm to introduce auxiliary tasks to AMR parsing, compared to multitask learning. From an empirical perspective, we propose a principled method to involve auxiliary tasks to boost AMR parsing. Extensive experiments show that our method achieves new state-of-the-art performance on different benchmarks especially in topology-related scores. Code and models are released at `https://github.com/PKUnlp-icler/ATP`.

SRW Thesis Proposals

16:15-17:45 (Quinault) *Chair: Daphne Ippolito*

Methods for Estimating and Improving Robustness of Language Models
Michal Stefanik 16:30-16:15 (Quinault)
Despite their outstanding performance, large language models (LLMs) suffer notorious flaws related to their preference for shallow textual relations over full semantic complexity of the problem. This proposal investigates a common denominator of this problem in their weak ability to generalise outside of the training domain. We survey diverse research directions providing estimations of model generalisation ability and find that incorporating some of these measures in the training objectives leads to enhanced distributional robustness of neural models. Based on these findings, we present future research directions enhancing the robustness of LLMs.

Retrieval-augmented Generation across Heterogeneous Knowledge
Wenhao Yu 16:30-16:45 (Quinault)
Retrieval-augmented generation (RAG) methods have been receiving increasing attention from the NLP community and achieved state-of-the-art performance on many NLP downstream tasks. Compared with conventional pre-trained generation models, RAG methods have remarkable advantages such as easy knowledge acquisition, strong scalability, and low training cost. Although existing RAG models have been applied to various knowledge-intensive NLP tasks, such as open-domain QA and dialogue systems, most of the work has focused on retrieving unstructured text documents from Wikipedia. In this paper, I first elaborate on the current obstacles to retrieving knowledge from a single-source homogeneous corpus. Then, I demonstrate evidence from both existing literature and my experiments, and provide multiple solutions on retrieval-augmented generation methods across heterogeneous knowledge.

Neural Retriever and Go Beyond: A Thesis Proposal
Man Luo 16:45-17:00 (Quinault)
Information Retriever (IR) aims to find the relevant documents (e.g. snippets, passages, and articles) to a given query at large scale. IR plays an important role in many tasks such as open domain question answering and dialogue systems, where external knowledge is needed. In the past, searching algorithms based on term matching have been widely used. Recently, neural-based algorithms (termed as neural retrievers) have gained more attention which can mitigate the limitations of traditional methods. Regardless of the success achieved by neural retrievers, they still face many challenges, e.g. suffering from a small amount of training data and failing to answer simple entity-centric questions. Furthermore, most of the existing neural retrievers are developed for pure-text query. This prevents them from handling multi-modality queries (i.e. the query is composed of textual description and images). This proposal has two goals. First, we introduce methods to address the abovementioned issues of neural retrievers from three angles, new model architectures, IR-oriented pretraining tasks, and generating large scale training data. Second, we identify the future research direction and propose potential corresponding solution.

Towards Gender Biased Language Classification: A Case Study with British English Archival Metadata Descriptions
Lucy Havens 16:45-17:00 (Quinault)
This thesis-in-progress summarizes the work completed and potential directions for a Ph.D. project researching the classification of gender biased language. Recognizing bias as inherent in language and thus inevitable in natural language processing systems, the project aims to make bias transparent. An interdisciplinary methodology is applied to define gender bias, annotate documents according to that definition, and train classification models on the annotated dataset to identify types of gender bias. Having created a gender biased language taxonomy and an annotated dataset, the project now moves towards the development of document classification models. There are several directions the classifier development could follow. The project would benefit from participation in the Student Research Workshop to discuss which

direction would add the most valuable contribution to computational linguistics.

Multimodal large language models for inclusive collaboration learning tasks
Armanda Lewis 17:00-17:15 (Quinault)
This PhD project leverages advancements in multimodal large language models to build an inclusive collaboration feedback loop, in order to facilitate the automated detection, modeling, and feedback for participants developing general collaboration skills. This topic is important given the role of collaboration as an essential 21st century skill, the potential to ground large language models within learning theory and real-world practice, and the expressive potential of transformer models to support equity and inclusion. We address some concerns of integrating advances in natural language processing into downstream tasks such as the learning analytics feedback loop.

Industry/Demo Poster Session

16:15-17:45 (Regency A & B)

Scalable and Robust Self-Learning for Skill Routing in Large-Scale Conversational AI Systems
Mohammad Kachuee, Jinseok Nam, Sarthak Ahuja, Jin-Myung Won and Sungjin Lee 16:15-17:45 (Regency A & B)
Skill routing is an important component in large-scale conversational systems. In contrast to traditional rule-based skill routing, state-of-the-art systems use a model-based approach to enable natural conversations. To provide supervision signal required to train such models, ideas such as human annotation, replication of a rule-based system, relabeling based on user paraphrases, and bandit-based learning were suggested. However, these approaches: (a) do not scale in terms of the number of skills and skill on-boarding, (b) require a very costly expert annotation/rule-design, (c) introduce risks in the user experience with each model update. In this paper, we present a scalable self-learning approach to explore routing alternatives without causing abrupt policy changes that break the user experience, learn from the user interaction, and incrementally improve the routing via frequent model refreshes. To enable such robust frequent model updates, we suggest a simple and effective approach that ensures controlled policy updates for individual domains, followed by an off-policy evaluation for making deployment decisions without any need for lengthy A/B experimentation. We conduct various offline and online A/B experiments on a commercial large-scale conversational system to demonstrate the effectiveness of the proposed method in real-world production settings.

AB/BA analysis: A framework for estimating keyword spotting recall improvement while maintaining audio privacy
Raphael Petegrosso, VasistaKrishna Baderdinnni, Thibaud Senechal and Benjamin Bullough 16:15-17:45 (Regency A & B)
Evaluation of keyword spotting (KWS) systems that detect keywords in speech is a challenging task under realistic privacy constraints. The KWS is designed to only collect data when the keyword is present, limiting the availability of hard samples that may contain false negatives, and preventing direct estimation of model recall from production data. Alternatively, complementary data collected from other sources may not be fully representative of the real application. In this work, we propose an evaluation technique which we call AB/BA analysis. Our framework evaluates a candidate KWS model B against a baseline model A, using cross-dataset offline decoding for relative recall estimation, without requiring negative examples. Moreover, we propose a formulation with assumptions that allow estimation of relative false positive rate between models with low variance even when the number of false positives is small. Finally, we propose to leverage machine-generated soft labels, in a technique we call Semi-Supervised AB/BA analysis, that improves the analysis time, privacy, and cost. Experiments with both simulation and real data show that AB/BA analysis is successful at measuring recall improvement in conjunction with the trade-off in relative false positive rate.

Temporal Generalization for Spoken Language Understanding
Judith Gaspers, Anoop Kumar, Greg Ver Steeg and Aram Galstyan 16:15-17:45 (Regency A & B)
Spoken Language Understanding (SLU) models in industry applications are usually trained offline on historic data, but have to perform well on incoming user requests after deployment. Since the application data is not available at training time, this is formally similar to the domain generalization problem, where domains correspond to different temporal segments of the data, and the goal is to build a model that performs well on unseen domains, e.g., upcoming data. In this paper, we explore different strategies for achieving good temporal generalization, including instance weighting, temporal fine-tuning, learning temporal features and building a temporally-invariant model. Our results on data of large-scale SLU systems show that temporal information can be leveraged to improve temporal generalization for SLU models.

An End-to-End Dialogue Summarization System for Sales Calls
Abedelkadir Asi, Song Wang, Roy Eisenstadt, Dean Geckt, Yarin Kuper, Yi Mao and Royi Ronen 16:15-17:45 (Regency A & B)
Summarizing sales calls is a routine task performed manually by salespeople. We present a production system which combines generative models fine-tuned for customer-agent setting, with a human-in-the-loop user experience for an interactive summary curation process. We address challenging aspects of dialogue summarization task in a real-world setting including long input dialogues, content validation, lack of labeled data and quality evaluation. We show how GPT-3 can be leveraged as an offline data labeler to handle training data scarcity and accommodate privacy constraints in an industrial setting. Experiments show significant improvements by our models in tackling the summarization and content validation tasks on public datasets.

Controlled Data Generation via Insertion Operations for NLU
Manoj Kumar, Yuval Merhav, Haidar Khan, Rahul Gupta, Anna Rumshisky and Wael Hamza 16:15-17:45 (Regency A & B)
Use of synthetic data is rapidly emerging as a realistic alternative to manually annotating live traffic for industry-scale model building. Manual data annotation is slow, expensive and not preferred for meeting customer privacy expectations. Further, commercial natural language applications are required to support continuously evolving features as well as newly added experiences. To address these requirements, we propose a targeted synthetic data generation technique by inserting tokens into a given semantic signature. The generated data are used as additional training samples in the tasks of intent classification and named entity recognition. We evaluate on a real-world voice assistant dataset, and using only 33% of the available training set, we achieve the same accuracy as training with all available data. Further, we analyze the effects of data generation across varied real-world applications and propose heuristics that improve the task performance further.

Easy and Efficient Transformer: Scalable Inference Solution For Large NLP Model
Gongzheng Li, Yadong Xi, Jingzhen Ding, Duan Wang, Ziyang Luo, Rongsheng Zhang, Bai Liu, Changjie Fan, Xiaoxi Mao and Zeng Zhao
16:15-17:45 (Regency A & B)
Recently, large-scale transformer-based models have been proven to be effective over various tasks across many domains. Nevertheless, applying them in industrial production requires tedious and heavy works to reduce inference costs. To fill such a gap, we introduce a scalable inference solution: **Easy and Efficient Transformer (EET)**, including a series of transformer inference optimization at the algorithm and implementation levels. First, we design highly optimized kernels for long inputs and large hidden sizes. Second, we propose a flexible CUDA memory manager to reduce the memory footprint when deploying a large model. Compared with the state-of-the-art transformer inference

library (Faster Transformer v4.0), EET can achieve an average of 1.40-4.20x speedup on the transformer decoder layer with an A100 GPU.

Efficient Semi-supervised Consistency Training for Natural Language Understanding
George Leung and Joshua Tan 16:15-17:45 (Regency A & B)
Manually labeled training data is expensive, noisy, and often scarce, such as when developing new features or localizing existing features for a new region. In cases where labeled data is limited but unlabeled data is abundant, semi-supervised learning methods such as consistency training can be used to improve model performance, by training models to output consistent predictions between original and augmented versions of unlabeled data. In this work, we explore different data augmentation methods for consistency training (CT) on Natural Language Understanding (NLU) domain classification (DC) in the limited labeled data regime. We explore three types of augmentation techniques (human paraphrasing, back-translation, and dropout) for unlabeled data and train DC models to jointly minimize both the supervised loss and the consistency loss on unlabeled data. Our results demonstrate that DC models trained with CT methods and dropout based augmentation on only 0.1% (2,998 instances) of labeled data with the remainder as unlabeled can achieve a top-1 relative accuracy reduction of 12.25% compared to fully supervised model trained with 100% of labeled data, outperforming fully supervised models trained on 10x that amount of labeled data. The dropout-based augmentation achieves similar performance compare to back-translation based augmentation with much less computational resources. This paves the way for applications of using large scale unlabeled data for semi-supervised learning in production NLU systems.

Distantly Supervised Aspect Clustering And Naming For E-Commerce Reviews
Prateek Sircar, Aniket Chakrabarti, Deepak Gupta and Anirban Majumdar 16:15-17:45 (Regency A & B)
Product aspect extraction from reviews is a critical task for e-commerce services to understand customer preferences and pain points. While aspect phrases extraction and sentiment analysis have received a lot of attention, clustering of aspect phrases and assigning human readable names to clusters in e-commerce reviews is an extremely important and challenging problem due to the scale of the reviews that makes human review infeasible. In this paper, we propose fully automated methods for clustering aspect words and generating human readable names for the clusters without any manually labeled data. We train transformer based sentence embeddings that are aware of unique e-commerce language characteristics (eg. incomplete sentences, spelling and grammar errors, vernacular etc.). We also train transformer based sequence to sequence models to generate human readable aspect names from clusters. Both the models are trained using heuristic based distant supervision. Additionally, the models are used to improve each other. Extensive empirical testing showed that the clustering model improves the Silhouette Score by 64% when compared to the state-of-the-art baseline and the aspect naming model achieves a high ROUGE-L score of 0.79.

CULG: Commercial Universal Language Generation
Haonan Li, Yameng Huang, Yeyun Gong, Jian Jiao, Ruofei Zhang, Timothy Baldwin and Nan Duan 16:15-17:45 (Regency A & B)
Pre-trained language models (PLMs) have dramatically improved performance for many natural language processing (NLP) tasks in domains such as finance and healthcare. However, the application of PLMs in the domain of commerce, especially marketing and advertising, remains less studied. In this work, we adapt pre-training methods to the domain of commerce, by proposing CULG, a large-scale commercial universal language generation model which is pre-trained on a corpus drawn from 10 markets across 7 languages. We propose 4 commercial generation tasks and a two-stage training strategy for pre-training, and demonstrate that the proposed strategy yields performance improvements on three generation tasks as compared to single-stage pre-training. Extensive experiments show that our model outperforms other models by a large margin on commercial generation tasks, and we conclude with a discussion on additional applications over other markets, languages, and tasks.

Constraining word alignments with posterior regularization for label transfer
Thomas Gueudre and Kevin Martin Jose 16:15-17:45 (Regency A & B)
Unsupervised word alignments offer a lightweight and interpretable method to transfer labels from high- to low-resource languages, as long as semantically related words have the same label across languages. But such an assumption is often not true in industrial NLP pipelines, where multilingual annotation guidelines are complex and deviate from semantic consistency due to various factors (such as annotation difficulty, conflicting ontology, upcoming feature launches etc.);

We address this difficulty by constraining the alignments models to remain consistent with both source and target annotation guidelines , leveraging posterior regularization and labeled examples. We illustrate the overall approach using IBM 2 (fast_align) as a base model, and report results on both internal and external annotated datasets. We measure consistent accuracy improvements on the MultiATIS++ dataset over AWESoME, a popular transformer-based alignment model, in the label projection task ($+2.7\%$ at word-level and $+15\%$ at sentence-level), and show how even a small amount of target language annotations help substantially.

Explaining the Effectiveness of Multi-Task Learning for Efficient Knowledge Extraction from Spine MRI Reports
Arijit Sehanobish, McCullen Sandora, Nabila Abraham, Jayashri Pawar, Danielle Torres, Anasuya Das, Murray Becker, Richard Herzog, Benjamin Odry and Ron Vianu 16:15-17:45 (Regency A & B)
Pretrained Transformer based models finetuned on domain specific corpora have changed the landscape of NLP. However, training or fine-tuning these models for individual tasks can be time consuming and resource intensive. Thus, a lot of current research is focused on using transformers for multi-task learning (Raffel et al., 2020) and how to group the tasks to help a multi-task model to learn effective representations that can be shared across tasks (Standley et al., 2020; Fifty et al., 2021) . In this work, we show that a single multi-tasking model can match the performance of task specific model when the task specific models show similar representations across all of their hidden layers and their gradients are aligned, i.e. their gradients follow the same direction. We hypothesize that the above observations explain the effectiveness of multi-task learning. We validate our observations on our internal radiologist-annotated datasets on the cervical and lumbar spine. Our method is simple and intuitive, and can be used in a wide range of NLP problems.

Asynchronous Convergence in Multi-Task Learning via Knowledge Distillation from Converged Tasks
Weiyi Lu, Sunny Rajagopalan, Priyanka Nigam, Jaspreet Singh, Xiaodi Sun, Yi Xu, Belinda Zeng and Trishul Chilimbi 16:15-17:45 (Regency A & B)
Multi-task learning (MTL) aims to solve multiple tasks jointly by sharing a base representation among them. This can lead to more efficient learning and better generalization, as compared to learning each task individually. However, one issue that often arises in MTL is the convergence speed between tasks varies due to differences in task difficulty, so it can be a challenge to simultaneously achieve the best performance on all tasks with a single model checkpoint. Various techniques have been proposed to address discrepancies in task convergence rate, including weighting the per-task losses and modifying task gradients. In this work, we propose a novel approach that avoids the problem of requiring all tasks to converge at the same rate, but rather allows for "asynchronous" convergence among the tasks where each task can converge on its own schedule. As our main contribution, we monitor per-task validation metrics and switch to a knowledge distillation loss once a task has converged instead of continuing to train on the true labels. This prevents the model from overfitting on converged tasks while it learns the remaining tasks. We evaluate the proposed method in two 5-task MTL setups consisting of internal e-commerce datasets. The results show that our method consistently outperforms existing loss weighting and gradient balancing approaches, achieving average improvements of 0.9% and 1.5% over the best performing baseline model in the two setups, respectively.

Augmenting Training Data for Massive Semantic Matching Models in Low-Traffic E-commerce Stores

Ashutosh Joshi, Shankar Vishwanath, Choon Hui Teo, Vaclav Petricek, Vishy Vishwanathan, Rahul Bhagat and Jonathan May 16:15-17:45 (Regency A & B)
Extreme multi-label classification (XMC) systems have been successfully applied in e-commerce (Shen et al., 2020; Dahiya et al., 2021) for retrieving products based on customer behavior. Such systems require large amounts of customer behavior data (e.g. queries, clicks, purchases) for training. However, behavioral data is limited in low-traffic e-commerce stores, impacting performance of these systems. In this paper, we present a technique that augments behavioral training data via query reformulation. We use the Aggregated Label eXtreme Multi-label Classification (AL-XMC) system (Shen et al., 2020) as an example semantic matching model and show via crowd-sourced human judgments that, when the training data is augmented through query reformulations, the quality of AL-XMC improves over a baseline that does not use query reformulation. We also show in online A/B tests that our method significantly improves business metrics for the AL-XMC model.

Retrieval Based Response Letter Generation For a Customer Care Setting
Biplob Biswas, Renhao Cui and Rajiv Ramnath 16:15-17:45 (Regency A & B)
Letter-like communications (such as email) are a major means of customer relationship management within customer-facing organizations. These communications are initiated on a channel by requests from customers and then responded to by the organization on the same channel. For decades, the job has almost entirely been conducted by human agents who attempt to provide the most appropriate reaction to the request. Rules have been made to standardize the overall customer service process and make sure the customers receive professional responses. Recent progress in natural language processing has made it possible to automate response generation. However, the diversity and open nature of customer queries and the lack of structured knowledge bases make this task even more challenging than typical task-oriented language generation tasks. Keeping those obstacles in mind, we propose a deep-learning based response letter generation framework that attempts to retrieve knowledge from historical responses and utilize it to generate an appropriate reply. Our model uses data augmentation to address the insufficiency of query-response pairs and employs a ranking mechanism to choose the best response from multiple potential options. We show that our technique outperforms the baselines by significant margins while producing consistent and informative responses.

Knowledge extraction from aeronautical messages (NOTAMs) with self-supervised language models for aircraft pilots
Alexandre Arnold, Fares Ernez, Catherine Kobus and Marion-Cécile Martin 16:15-17:45 (Regency A & B)
During their pre-flight briefings, aircraft pilots must analyse a long list of NoTAMs (NOtice To AirMen) indicating potential hazards along the flight route, sometimes up to pages for long-haul flights. NOTAM free-text fields typically have a very special phrasing, with lots of acronyms and domain-specific vocabulary, which makes it differ significantly from standard English. In this paper, we pretrain language models derived from BERT on circa 1 million unlabeled NOTAMs and reuse the learnt representations on three downstream tasks valuable for pilots: criticality prediction, named entity recognition and translation into a structured language called Airlang. This self-supervised approach, where smaller amounts of labeled data are enough for task-specific fine-tuning, is well suited in the aeronautical context since expert annotations are expensive and time-consuming. We present evaluation scores across the tasks showing a high potential for an operational usability of such models (by pilots, airlines or service providers), which is a first to the best of our knowledge.

Intent Discovery for Enterprise Virtual Assistants: Applications of Utterance Embedding and Clustering to Intent Mining
Minhua Chen, Badrinath Jayakumar, Michael Johnston, S. Eman Mahmoodi and Daniel Pressel 16:15-17:45 (Regency A & B)
A key challenge in the creation and refinement of virtual assistants is the ability to mine unlabeled utterance data to discover common intents. We develop an approach to this problem that combines large-scale pre-training and multi-task learning to derive a semantic embedding that can be leveraged to identify clusters of utterances that correspond to unhandled intents. An utterance encoder is first trained with a language modeling objective and subsequently adapted to predict intent labels from a large collection of cross-domain enterprise virtual assistant data using a multi-task cosine softmax loss. Experimental evaluation shows significant advantages for this multi-step pre-training approach, with large gains in downstream clustering accuracy on new applications compared to standard sentence embedding approaches. The approach has been incorporated into an interactive discovery tool that enables visualization and exploration of intents by system analysts and builders.

Lightweight Transformers for Conversational AI
Daniel Pressel, Wenshuo Liu, Michael Johnston and Minhua Chen 16:15-17:45 (Regency A & B)
To understand how training on conversational language impacts performance of pre-trained models on downstream dialogue tasks, we build compact Transformer-based Language Models from scratch on several large corpora of conversational data. We compare the performance and characteristics of these models against BERT and other strong baselines on dialogue probing tasks. Commercial dialogue systems typically require a small footprint and fast execution time, but recent trends are in the other direction, with an ever-increasing number of parameters, resulting in difficulties in model deployment. We focus instead on training fast, lightweight models that excel at natural language understanding (NLU) and can replace existing lower-capacity conversational AI models with similar size and speed. In the process, we develop a simple but unique curriculum-based approach that moves from general-purpose to dialogue-targeted both in terms of data and objective. Our resultant models have around 1/3 the number of parameters of BERT-base and produce better representations for a wide array of intent detection datasets using linear and Mutual-Information probing techniques. Additionally, the models can be easily fine-tuned on a single consumer GPU card and deployed in near real-time production environments.

NER-MQMRC: Formulating Named Entity Recognition as Multi Question Machine Reading Comprehension
Anubhav Shrimal, Avi Jain, Kartik Mehta and Promod Yenigalla 16:15-17:45 (Regency A & B)
NER has been traditionally formulated as a sequence labeling task. However, there has been recent trend in posing NER as a machine reading comprehension task (Wang et al., 2020; Mengge et al., 2020), where entity name (or other information) is considered as a question, text as the context and entity value in text as answer snippet. These works consider MRC based on a single question (entity) at a time. We propose posing NER as a multi-question MRC task, where multiple questions (one question per entity) are considered at the same time for a single text. We propose a novel BERT-based multi-question MRC (NER-MQMRC) architecture for this formulation. NER-MQMRC architecture considers all entities as input to BERT for learning token embeddings with self-attention and leverages BERT-based entity representation for further improving these token embeddings for NER task. Evaluation on three NER datasets show that our proposed architecture leads to average 2.5 times faster training and 2.3 times faster inference as compared to NER-SQMRC framework based models by considering all entities together in a single pass. Further, we show that our model performance does not degrade compared to single-question based MRC (NER-SQMRC) (Devlin et al., 2019) leading to F1 gain of +0.41%, +0.32% and +0.27% for AE-Pub, Ecommerce5PT and Twitter datasets respectively. We propose this architecture primarily to solve large scale e-commerce attribute (or entity) extraction from unstructured text of a magnitude of 50k+ attributes to be extracted on a scalable production environment with high performance and optimised training and inference runtimes.

What Do Users Care About? Detecting Actionable Insights from User Feedback
Kasturi Bhattacharjee, Rashmi Gangadharaiah, Kathleen McKeown and Dan Roth 16:15-17:45 (Regency A & B)
Users often leave feedback on a myriad of aspects of a product which, if leveraged successfully, can help yield useful insights that can lead to further improvements down the line. Detecting actionable insights can be challenging owing to large amounts of data as well as the absence of labels in real-world scenarios. In this work, we present an aggregation and graph-based ranking strategy for unsupervised detection of these insights from real-world, noisy, user-generated feedback. Our proposed approach significantly outperforms strong baselines on two

real-world user feedback datasets and one academic dataset.

Developing a Production System for Purpose of Call Detection in Business Phone Conversations
Elena Khasanova, Pooja Hiranandani, Shayna Gardiner, Cheng Chen, Simon Corston-Oliver and Xue-Yong Fu 16:15-17:45 (Regency A & B)
For agents at a contact centre receiving calls, the most important piece of information is the reason for a given call. An agent cannot provide support on a call if they do not know why a customer is calling. In this paper we describe our implementation of a commercial system to detect Purpose of Call statements in English business call transcripts in real time. We present a detailed analysis of types of Purpose of Call statements and language patterns related to them, discuss an approach to collect rich training data by bootstrapping from a set of rules to a neural model, and describe a hybrid model which consists of a transformer-based classifier and a set of rules by leveraging insights from the analysis of call transcripts. The model achieved 88.6 F1 on average in various types of business calls when tested on real life data and has low inference time. We reflect on the challenges and design decisions when developing and deploying the system.

Adversarial Text Normalization
Joanna Bitton, Maya Pavlova and Ivan Evtimov 16:15-17:45 (Regency A & B)
Text-based adversarial attacks are becoming more commonplace and accessible to general internet users. As these attacks proliferate, the need to address the gap in model robustness becomes imminent. While retraining on adversarial data may increase performance, there remains an additional class of character-level attacks on which these models falter. Additionally, the process to retrain a model is time and resource intensive, creating a need for a lightweight, reusable defense. In this work, we propose the Adversarial Text Normalizer, a novel method that restores baseline performance on attacked content with low computational overhead. We evaluate the efficacy of the normalizer on two problem areas prone to adversarial attacks, i.e. Hate Speech and Natural Language Inference. We find that text normalization provides a task-agnostic defense against character-level attacks that can be implemented supplementary to adversarial retraining solutions, which are more suited for semantic alterations.

Constraint-based Multi-hop Question Answering with Knowledge Graph
Sayantan Mitra, Roshni Ramnani and Shubhashis Sengupta 16:15-17:45 (Regency A & B)
The objective of a Question-Answering system over Knowledge Graph (KGQA) is to respond to natural language queries presented over the KG. A complex question answering system typically addresses one of the two categories of complexity: questions with constraints and questions involving multiple hops of relations. Most of the previous works have addressed these complexities separately. Multi-hop KGQA necessitates reasoning across numerous edges of the KG in order to arrive at the correct answer. Because KGs are frequently sparse, multi-hop KGQA presents extra complications. Recent works have developed KG embedding approaches to reduce KG sparsity by performing missing link prediction. In this paper, we tried to address multi-hop constrained-based queries using KG embeddings to generate more flexible query graphs. Empirical results indicate that the proposed methodology produces state-of-the-art outcomes on three KGQA datasets.

Fast Bilingual Grapheme-To-Phoneme Conversion
Hwa-Yeon Kim, Jong-Hwan Kim and Jae-Min Kim 16:15-17:45 (Regency A & B)
Autoregressive transformer (ART)-based grapheme-to-phoneme (G2P) models have been proposed for bi/multilingual text-to-speech systems. Although they have achieved great success, they suffer from high inference latency in real-time industrial applications, especially processing long sentence. In this paper, we propose a fast and high-performance bilingual G2P model. For fast and exact decoding, we used a non-autoregressive structured transformer-based architecture and data augmentation for predicting output length. Our model achieved better performance than that of the previous autoregressive model and about 2700% faster inference speed.

Knowledge Extraction From Texts Based on Wikidata
Anastasia Shimorina, Johannes Heinecke and Frédéric Herledan 16:15-17:45 (Regency A & B)
This paper presents an effort within our company of developing knowledge extraction pipeline for English, which can be further used for constructing an entreprise-specific knowledge base. We present a system consisting of entity detection and linking, coreference resolution, and relation extraction based on the Wikidata schema. We highlight existing challenges of knowledge extraction by evaluating the deployed pipeline on real-world data. We also make available a database, which can serve as a new resource for sentential relation extraction, and we underline the importance of having balanced data for training classification models.

AIT-QA: Question Answering Dataset over Complex Tables in the Airline Industry
Yannis Katsis, Saneem Ahmed Chemmengath, Vishwajeet Kumar, Samarth Bharadwaj, Mustafa Canim, Michael Glass, Alfio Gliozzo, Feifei Pan, Jaydeep Sen, Karthik Sankaranarayanan and Soumen Chakrabarti 16:15-17:45 (Regency A & B)
Table Question Answering (Table QA) systems have been shown to be highly accurate when trained and tested on open-domain datasets built on top of Wikipedia tables. However, it is not clear whether their performance remains the same when applied to domain-specific scientific and business documents, encountered in industrial settings, which exhibit some unique characteristics: (a) they contain tables with a much more complex layout than Wikipedia tables (including hierarchical row and column headers), (b) they contain domain-specific terms, and (c) they are typically not accompanied by domain-specific labeled data that can be used to train Table QA models. To understand the performance of Table QA approaches in this setting, we introduce AIT-QA; a domain-specific Table QA test dataset. While focusing on the airline industry, AIT-QA reflects the challenges that domain-specific documents pose to Table QA, outlined above. In this work, we describe the creation of the dataset and report zero-shot experimental results of three SOTA Table QA methods. The results clearly expose the limitations of current methods with a best accuracy of just 51.8%. We also present pragmatic table pre-processing steps to pivot and project complex tables into a layout suitable for the SOTA Table QA models. Finally, we provide data-driven insights on how different aspects of this setting (including hierarchical headers, domain-specific terminology, and paraphrasing) affect Table QA methods, in order to help the community develop improved methods for domain-specific Table QA.

Parameter-efficient Continual Learning Framework in Industrial Real-time Text Classification System
Tao Zhu, Zhe Zhao, Weijie Liu, Jiachi Liu, Yiren Chen, Weiquan Mao, Haoyan Liu, Kunbo Ding, Yudong Li and Xuefeng Yang 16:15-17:45 (Regency A & B)
Catastrophic forgetting is a challenge for model deployment in industrial real-time systems, which requires the model to quickly master a new task without forgetting the old one. Continual learning aims to solve this problem; however, it usually updates all the model parameters, resulting in extensive training times and the inability to deploy quickly. To address this challenge, we propose a parameter-efficient continual learning framework, in which efficient parameters are selected through an offline parameter selection strategy and then trained using an online regularization method. In our framework, only a few parameters need to be updated, which not only alleviates catastrophic forgetting, but also allows the model to be saved with the changed parameters instead of all parameters. Extensive experiments are conducted to examine the effectiveness of our proposal. We believe this paper will provide useful insights and experiences on developing deep learning-based online real-time systems.

Fast and Light-Weight Answer Text Retrieval in Dialogue Systems

Hui Wan, Siva Sankalp Patel, J William Murdock, Saloni Potdar and Sachindra Joshi 16:15-17:45 (Regency A & B)
Dialogue systems can benefit from being able to search through a corpus of text to find information relevant to user requests, especially when encountering a request for which no manually curated response is available. The state-of-the-art technology for neural dense retrieval or re-ranking involves deep learning models with hundreds of millions of parameters. However, it is difficult and expensive to get such models to operate at an industrial scale, especially for cloud services that often need to support a big number of individually customized dialogue systems, each with its own text corpus. We report our work on enabling advanced neural dense retrieval systems to operate effectively at scale on relatively inexpensive hardware. We compare with leading alternative industrial solutions and show that we can provide a solution that is effective, fast, and cost-efficient.

BLINK with Elasticsearch for Efficient Entity Linking in Business Conversations
Md Tahmid Rahman Laskar, Cheng Chen, Aliaksandr Martsinovich, Jonathan Johnston, Xue-Yong Fu, Shashi Bhushan TN and Simon Corston-Oliver 16:15-17:45 (Regency A & B)
An Entity Linking system aligns the textual mentions of entities in a text to their corresponding entries in a knowledge base. However, deploying a neural entity linking system for efficient real-time inference in production environments is a challenging task. In this work, we present a neural entity linking system that connects the product and organization type entities in business conversations to their corresponding Wikipedia and Wikidata entries. The proposed system leverages Elasticsearch to ensure inference efficiency when deployed in a resource limited cloud machine, and obtains significant improvements in terms of inference speed and memory consumption while retaining high accuracy.

Q2R: A Query-to-Resolution System for Natural-Language Queries
Shiau Hong Lim and Laura Wynter 16:15-17:45 (Regency A & B)
We present a system for document retrieval that combines direct classification with standard content-based retrieval approaches to significantly improve the relevance of the retrieved documents. Our system exploits the availability of an imperfect but sizable amount of labeled data from past queries. For domains such as technical support, the proposed approach enhances the system's ability to retrieve documents that are otherwise ranked very low based on content alone. The system is easy to implement and can make use of existing text ranking methods, augmenting them through the novel Q2R orchestration framework. Q2R has been extensively tested and is in use at IBM.

Identifying Corporate Credit Risk Sentiments from Financial News
Noujoud Ahbali, Xinyuan Liu, Albert Aristotle Nanda, Jamie Stark, Ashit Talukder and Rupinder Paul Khandpur 16:15-17:45 (Regency A & B)
Credit risk management is one central practice for financial institutions, and such practice helps them measure and understand the inherent risk within their portfolios. Historically, firms relied on the assessment of default probabilities and used the press as one tool to gather insights on the latest credit event developments of an entity. However, due to the deluge of the current news coverage for companies, analyzing news manually by financial experts is considered a highly laborious task. To this end, we propose a novel deep learning-powered approach to automate news analysis and credit adverse events detection to score the credit sentiment associated with a company. This paper showcases a complete system that leverages news extraction and data enrichment with targeted sentiment entity recognition to detect companies and text classification to identify credit events. We developed a custom scoring mechanism to provide the company's credit sentiment score (CSS^{TM}) based on these detected events. Additionally, using case studies, we illustrate how this score helps understand the company's credit profile and discriminates between defaulters and non-defaulters.

textless-lib: a Library for Textless Spoken Language Processing
Eugene Kharitonov, Jade Copet, Kushal Lakhotia, Tu Anh Nguyen, Paden Tomasello, Ann Lee, Ali Elkahky, Wei-Ning Hsu, Abdelrahman Mohamed, Emmanuel Dupoux and Yossi Adi 16:15-17:45 (Regency A & B)
Textless spoken language processing is an exciting area of research that promises to extend applicability of the standard NLP toolset onto spoken language and languages with few or no textual resources. Here, we introduce textless-lib, a PyTorch-based library aimed to facilitate research in the area. We describe the building blocks that the library provides and demonstrate its usability by discuss three different use-case examples: (i) speaker probing, (ii) speech resynthesis and compression, and (iii) speech continuation. We believe that `textless-lib` substantially simplifies research the textless setting and will be handful not only for speech researchers but also for the NLP community at large.

Web-based Annotation Interface for Derivational Morphology
Lukáš Kyjánek 16:15-17:45 (Regency A & B)
The paper presents a visual interface for manual annotation of language resources for derivational morphology. The interface is web-based and created using relatively simple programming techniques, and yet it rapidly facilitates and speeds up the annotation process, especially in languages with rich derivational morphology. As such, it can reduce the cost of the process. After introducing manual annotation tasks in derivational morphology, the paper describes the new visual interface and a case study that compares the current annotation method to the annotation using the interface. In addition, it also demonstrates the opportunity to use the interface for manual annotation of syntactic trees. The source codes are freely available under the MIT License on GitHub.

TurkishDelightNLP: A Neural Turkish NLP Toolkit
Huseyin Alecakir, Necva Bölücü and Burcu Can 16:15-17:45 (Regency A & B)
We introduce a neural Turkish NLP toolkit called TurkishDelightNLP that performs computational linguistic analyses from morphological level to semantic level that involves tasks such as stemming, morphological segmentation, morphological tagging, part-of-speech tagging, dependency parsing, and semantic parsing, as well as high-level NLP tasks such as named entity recognition. We publicly share the open-source Turkish NLP toolkit through a web interface that allows an input text to be analysed in real-time, as well as the open source implementation of the components provided in the toolkit, an API, and several annotated datasets such as word similarity test set to evaluate word embeddings and UCCA-based semantic annotation in Turkish. This will be the first open-source Turkish NLP toolkit that involves a range of NLP tasks in all levels. We believe that it will be useful for other researchers in Turkish NLP and will be also beneficial for other high-level NLP tasks in Turkish.

ZS4IE: A toolkit for Zero-Shot Information Extraction with simple Verbalizations
Oscar Sainz, Haoling Qiu, Oier Lopez De Lacalle, Eneko Agirre and Bonan Min 16:15-17:45 (Regency A & B)
The current workflow for Information Extraction (IE) analysts involves the definition of the entities/relations of interest and a training corpus with annotated examples. In this demonstration we introduce a new workflow where the analyst directly verbalizes the entities/relations, which are then used by a Textual Entailment model to perform zero-shot IE. We present the design and implementation of a toolkit with a user interface, as well as experiments on four IE tasks that show that the system achieves very good performance at zero-shot learning using only 5–15 minutes per type of a user's effort. Our demonstration system is open-sourced at https://github.com/BBN-E/ZS4IE. A demonstration video is available at https://vimeo.com/676138340.

Flowstorm: Open-Source Platform with Hybrid Dialogue Architecture

Jan Pichl, Petr Marek, Jakub Konrád, Petr Lorenc, Ondrej Kobza, Tomáš Zajíček and Jan Šedivý 16:15-17:45 (Regency A & B)

This paper presents a conversational AI platform called Flowstorm. Flowstorm is an open-source SaaS project suitable for creating, running, and analyzing conversational applications. Thanks to the fast and fully automated build process, the dialogues created within the platform can be executed in seconds. Furthermore, we propose a novel dialogue architecture that uses a combination of tree structures with generative models. The tree structures are also used for training NLU models suitable for specific dialogue scenarios. However, the generative models are globally used across applications and extend the functionality of the dialogue trees. Moreover, the platform functionality benefits from out-of-the-box components, such as the one responsible for extracting data from utterances or working with crawled data. Additionally, it can be extended using a custom code directly in the platform. One of the essential features of the platform is the possibility to reuse the created assets across applications. There is a library of prepared assets where each developer can contribute. All of the features are available through a user-friendly visual editor.

Contrastive Explanations of Text Classifiers as a Service

Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, Navid Nobani and Andrea Seveso 16:15-17:45 (Regency A & B)

The recent growth of black-box machine-learning methods in data analysis has increased the demand for explanation methods and tools to understand their behaviour and assist human-ML model cooperation. In this paper, we demonstrate ContrXT, a novel approach that uses natural language explanations to help users to comprehend how a back-box model works. ContrXT provides time contrastive (t-contrast) explanations by computing the differences in the classification logic of two different trained models and then reasoning on their symbolic representations through Binary Decision Diagrams. ContrXT is publicly available at ContrXT.ai as a python pip package.

RESIN-11: Schema-guided Event Prediction for 11 Newsworthy Scenarios

Xinya Du, Zixuan Zhang, Sha Li, Pengfei Yu, Hongwei Wang, Tuan Lai, Xudong Lin, Ziqi Wang, Iris Liu, Ben Zhou, Haoyang Wen, Manling Li, Darryl Hannan, Jie Lei, Hyounghun Kim, Rotem Dror, Haoyu Wang, Michael Regan, Qi Zeng, Qing Lyu, Charles Yu, Carl Edwards, Xiaomeng Jin, Yizhu Jiao, Ghazaleh Kazeminejad, Zhenhailong Wang, Chris Callison-Burch, Mohit Bansal, Carl Vondrick, Jiawei Han, Dan Roth, Shih-Fu Chang, Martha Palmer and Heng Ji 16:15-17:45 (Regency A & B)

We introduce RESIN-11, a new schema-guided event extraction&prediction framework that can be applied to a large variety of newsworthy scenarios. The framework consists of two parts: (1) an open-domain end-to-end multimedia multilingual information extraction system with weak-supervision and zero-shot learningbased techniques. (2) schema matching and schema-guided event prediction based on our curated schema library. We build a demo website based on our dockerized system and schema library publicly available for installation (https://github.com/RESIN-KAIROS/RESIN-11). We also include a video demonstrating the system.

A Human-machine Interface for Few-shot Rule Synthesis for Information Extraction

Robert Vacareanu, George C.G. Barbosa, Enrique Noriega-Atala, Gus Hahn-Powell, Rebecca Sharp, Marco A. Valenzuela-Escárcega and Mihai Surdeanu 16:15-17:45 (Regency A & B)

We propose a system that assists a user in constructing transparent information extraction models, consisting of patterns (or rules) written in a declarative language, through program synthesis. Users of our system can specify their requirements through the use of examples, which are collected with a search interface. The rule-synthesis system proposes rule candidates and the results of applying them on a textual corpus; the user has the option to accept the candidate, request another option, or adjust the examples provided to the system. Through an interactive evaluation, we show that our approach generates high-precision rules even in a 1-shot setting. On a second evaluation on a widely-used relation extraction dataset (TACRED), our method generates rules that outperform considerably manually written patterns. Our code, demo, and documentation is available at https://clulab.github.io/odinsynth.

SETSum: Summarization and Visualization of Student Evaluations of Teaching

Yinuo Hu, Shiyue Zhang, Viji Sathy, Abigail Panter and Mohit Bansal 16:15-17:45 (Regency A & B)

Student Evaluations of Teaching (SETs) are widely used in colleges and universities. Typically SET results are summarized for instructors in a static PDF report. The report often includes summary statistics for quantitative ratings and an unsorted list of open-ended student comments. The lack of organization and summarization of the raw comments hinders those interpreting the reports from fully utilizing informative feedback, making accurate inferences, and designing appropriate instructional improvements. In this work, we introduce a novel system, SETSUM, that leverages sentiment analysis, aspect extraction, summarization, and visualization techniques to provide organized illustrations of SET findings to instructors and other reviewers. Ten university professors from diverse departments serve as evaluators of the system and all agree that SETSUM help them interpret SET results more efficiently; and 6 out of 10 instructors prefer our system over the standard static PDF report (while the remaining 4 would like to have both). This demonstrates that our work holds the potential of reforming the SET reporting conventions in the future.

Towards Open-Domain Topic Classification

Hantian Ding, Jinrui Yang, Yuqian Deng, Hongming Zhang and Dan Roth 16:15-17:45 (Regency A & B)

We introduce an open-domain topic classification system that accepts user-defined taxonomy in real time. Users will be able to classify a text snippet with respect to any candidate labels they want, and get instant response from our web interface. To obtain such flexibility, we build the backend model in a zero-shot way. By training on a new dataset constructed from Wikipedia, our label-aware text classifier can effectively utilize implicit knowledge in the pretrained language model to handle labels it has never seen before. We evaluate our model across four datasets from various domains with different label sets. Experiments show that the model significantly improves over existing zero-shot baselines in open-domain scenarios, and performs competitively with weakly-supervised models trained on in-domain data.

SentSpace: Large-Scale Benchmarking and Evaluation of Text using Cognitively Motivated Lexical, Syntactic, and Semantic Features

Greta Tuckute, Aalok Sathe, Mingye Wang, Harley Yoder, Cory Shain and Evelina Fedorenko 16:15-17:45 (Regency A & B)

SentSpace is a modular framework for streamlined evaluation of text. SentSpace characterizes textual input using diverse lexical, syntactic, and semantic features derived from corpora and psycholinguistic experiments. Core sentence features fall into three primary feature spaces: 1) Lexical, 2) Contextual, and 3) Embeddings. To aid in the analysis of computed features, SentSpace provides a web interface for interactive visualization and comparison with text from large corpora. The modular design of SentSpace allows researchers to easily integrate their own feature computation into the pipeline while benefiting from a common framework for evaluation and visualization. In this manuscript we will describe the design of SentSpace, its core feature spaces, and demonstrate an example use case by comparing human-written and machine-generated (GPT2-XL) sentences to each other. We find that while GPT2-XL-generated text appears fluent at the surface level, psycholinguistic norms and measures of syntactic processing reveal key differences between text produced by humans and machines. Thus, SentSpace provides a broad set of cognitively motivated linguistic features for evaluation of text within natural language processing, cognitive science, as well as the social sciences.

PaddleSpeech: An Easy-to-Use All-in-One Speech Toolkit

Hui Zhang, Tian Yuan, Junkun Chen, Xintong Li, Renjie Zheng, Yuxin Huang, Xiaojie Chen, Enlei Gong, Zeyu Chen, Xiaoguang Hu, Dianhai

Yu, Yanjun Ma and Liang Huang 16:15-17:45 (Regency A & B)
PaddleSpeech is an open-source all-in-one speech toolkit. It aims at facilitating the development and research of speech processing technologies by providing an easy-to-use command-line interface and a simple code structure. This paper describes the design philosophy and core architecture of PaddleSpeech to support several essential speech-to-text and text-to-speech tasks. PaddleSpeech achieves competitive or state-of-the-art performance on various speech datasets and implements the most popular methods. It also provides recipes and pretrained models to quickly reproduce the experimental results in this paper. PaddleSpeech is publicly avaiable at https://github.com/PaddlePaddle/PaddleSpeech.

DadmaTools: Natural Language Processing Toolkit for Persian Language
*Romina Etezadi, Mohammad Karrabi, Najmeh Zare, Mohamad Bagher Sajadi and Mohammad Taher Pilehvar*16:15-17:45 (Regency A & B)
We introduce DadmaTools, an open-source Python Natural Language Processing toolkit for the Persian language. The toolkit is a neural pipeline based on spaCy for several text processing tasks, including normalization, tokenization, lemmatization, part-of-speech, dependency parsing, constituency parsing, chunking, and ezafe detecting. DadmaTools relies on fine-tuning of ParsBERT using the PerDT dataset for most of the tasks. Dataset module and embedding module are included in DadmaTools that support different Persian datasets, embeddings, and commonly used functions for them. Our evaluations show that DadmaTools can attain state-of-the-art performance on multiple NLP tasks. The source code is freely available at https://github.com/Dadmatech/DadmaTools.

FAMIE: A Fast Active Learning Framework for Multilingual Information Extraction
Minh Van Nguyen, Nghia Trung Ngo, Bonan Min and Thien Huu Nguyen 16:15-17:45 (Regency A & B)
This paper presents FAMIE, a comprehensive and efficient active learning (AL) toolkit for multilingual information extraction. FAMIE is designed to address a fundamental problem in existing AL frameworks where annotators need to wait for a long time between annotation batches due to the time-consuming nature of model training and data selection at each AL iteration. This hinders the engagement, productivity, and efficiency of annotators. Based on the idea of using a small proxy network for fast data selection, we introduce a novel knowledge distillation mechanism to synchronize the proxy network with the main large model (i.e., BERT-based) to ensure the appropriateness of the selected annotation examples for the main model. Our AL framework can support multiple languages. The experiments demonstrate the advantages of FAMIE in terms of competitive performance and time efficiency for sequence labeling with AL. We publicly release our code (https://github.com/nlp-uoregon/famie) and demo website (http://nlp.uoregon.edu:9000/). A demo video for FAMIE is provided at: https://youtu.be/I2i8n_jAyrY

Main Conference: Wednesday, July 13, 2022

Session 7 - 08:00-09:00

Interpretability and Analysis of Models for NLP 2

08:00-09:00 (Columbia A)

When a sentence does not introduce a discourse entity, Transformer-based models still sometimes refer to it
Sebastian Schuster and Tal Linzen 08:00-08:15 (Columbia A)
Understanding longer narratives or participating in conversations requires tracking of discourse entities that have been mentioned. Indefinite noun phrases (NPs), such as 'a dog', frequently introduce discourse entities but this behavior is modulated by sentential operators such as negation. For example, 'a dog' in 'Arthur doesn't own a dog' does not introduce a discourse entity due to the presence of negation. In this work, we adapt the psycholinguistic assessment of language models paradigm to higher-level linguistic phenomena and introduce an English evaluation suite that targets the knowledge of the interactions between sentential operators and indefinite NPs. We use this evaluation suite for a fine-grained investigation of the entity tracking abilities of the Transformer-based models GPT-2 and GPT-3. We find that while the models are to a certain extent sensitive to the interactions we investigate, they are all challenged by the presence of multiple NPs and their behavior is not systematic, which suggests that even models at the scale of GPT-3 do not fully acquire basic entity tracking abilities.

Analyzing Encoded Concepts in Transformer Language Models
Hassan Sajjad, Nadir Durrani, Fahim Dalvi, Firoj Alam, Abdul Rafae Khan and Jia Xu 08:15-08:30 (Columbia A)
We propose a novel framework ConceptX, to analyze how latent concepts are encoded in representations learned within pre-trained lan- guage models. It uses clustering to discover the encoded concepts and explains them by aligning with a large set of human-defined concepts. Our analysis on seven transformer language models reveal interesting insights: i) the latent space within the learned representations overlap with different linguistic concepts to a varying degree, ii) the lower layers in the model are dominated by lexical concepts (e.g., affixation) and linguistic ontologies (e.g. Word-Net), whereas the core-linguistic concepts (e.g., morphology, syntactic relations) are better represented in the middle and higher layers, iii) some encoded concepts are multi-faceted and cannot be adequately explained using the existing human-defined concepts.

Probing via Prompting
Jiaoda Li, Ryan Cotterell and Mrinmaya Sachan 08:30-08:45 (Columbia A)
Probing is a popular approach to understand what linguistic information is contained in the representations of pre-trained language models. However, the mechanism of selecting the probe model has recently been subject to intense debate, as it is not clear if the probes are merely extracting information or modelling the linguistic property themselves. To address this challenge, this paper introduces a novel model-free approach to probing via prompting, which formulates probing as a prompting task. We conduct experiments on five probing tasks and show that PP is comparable or better at extracting information than diagnostic probes while learning much less on its own. We further combine the probing via prompting approach with pruning to analyze where the model stores the linguistic information in its architecture. Finally, we apply the probing via prompting approach to examine the usefulness of a linguistic property for pre-training by removing the heads that are essential to it and evaluating the resulting model's performance on language modeling.

GlobEnc: Quantifying Global Token Attribution by Incorporating the Whole Encoder Layer in Transformers
Ali Modarressi, Mohsen Fayyaz, Yadollah Yaghoobzadeh and Mohammad Taher Pilehvar 08:45-09:00 (Columbia A)
There has been a growing interest in interpreting the underlying dynamics of Transformers. While self-attention patterns were initially deemed as the primary option, recent studies have shown that integrating other components can yield more accurate explanations. This paper introduces a novel token attribution analysis method that incorporates all the components in the encoder block and aggregates this throughout layers. Through extensive quantitative and qualitative experiments, we demonstrate that our method can produce faithful and meaningful global token attributions. Our experiments reveal that incorporating almost every encoder component results in increasingly more accurate analysis in both local (single layer) and global (the whole model) settings. Our global attribution analysis significantly outperforms previous methods on various tasks regarding correlation with gradient-based saliency scores. Our code is freely available at https://github.com/mohsenfayyaz/GlobEnc.

Summarization

08:00-09:00 (Columbia C)

From spoken dialogue to formal summary: An utterance rewriting for dialogue summarization
Yue Fang, Hainan Zhang, Hongshen Chen, Zhuoye Ding, Bo Long, Yanyan Lan and Yanquan Zhou 08:00-08:15 (Columbia C)
Due to the dialogue characteristics of unstructured contexts and multi-parties with first-person perspective, many successful text summarization works have failed when dealing with dialogue summarization. In dialogue summarization task, the input dialogue is usually spoken style with ellipsis and co-references but the output summaries are more formal and complete. Therefore, the dialogue summarization model should be able to complete the ellipsis content and co-reference information and then produce a suitable summary accordingly. However, the current state-of-the-art models pay more attention on the topic or structure of summary, rather than the consistency of dialogue summary with its input dialogue context, which may suffer from the personal and logical inconsistency problem. In this paper, we propose a new model, named ReWriteSum, to tackle this problem. Firstly, an utterance rewriter is conducted to complete the ellipsis content of dialogue content and then obtain the rewriting utterances. Then, the co-reference data augmentation mechanism is utilized to replace the referential person name with its specific name to enhance the personal information. Finally, the rewriting utterances and the co-reference replacement data are used in the standard BART model. Experimental results on both SAMSum and DialSum datasets show that our ReWriteSum significantly outperforms baseline models, in terms of both metric-based and human evaluations. Further analysis on multi-speakers also shows that ReWriteSum can obtain relatively higher improvement with more speakers, validating the correctness and property of ReWriteSum.

Domain-Oriented Prefix-Tuning: Towards Efficient and Generalizable Fine-tuning for Zero-Shot Dialogue Summarization

Lulu Zhao, Fujia Zheng, Weihao Zeng, Keqing He, Weiran Xu, Huixing Jiang, Wei Wu and Yanan Wu 08:15-08:30 (Columbia C)
The most advanced abstractive dialogue summarizers lack generalization ability on new domains and the existing researches for domain adaptation in summarization generally rely on large-scale pre-trainings. To explore the lightweight fine-tuning methods for domain adaptation of dialogue summarization, in this paper, we propose an efficient and generalizable Domain-Oriented Prefix-tuning model, which utilizes a domain word initialized prefix module to alleviate domain entanglement and adopts discrete prompts to guide the model to focus on key contents of dialogues and enhance model generalization. We conduct zero-shot experiments and build domain adaptation benchmarks on two multi-domain dialogue summarization datasets, TODSum and QMSum. Adequate experiments and qualitative analysis prove the effectiveness of our methods.

DialSummEval: Revisiting Summarization Evaluation for Dialogues
Mingqi Gao and Xiaojun Wan 08:30-08:45 (Columbia C)
Dialogue summarization is receiving increasing attention from researchers due to its extraordinary difficulty and unique application value. We observe that current dialogue summarization models have flaws that may not be well exposed by frequently used metrics such as ROUGE. In our paper, we re-evaluate 18 categories of metrics in terms of four dimensions: coherence, consistency, fluency and relevance, as well as a unified human evaluation of various models for the first time. Some noteworthy trends which are different from the conventional summarization tasks are identified. We will release DialSummEval, a multi-faceted dataset of human judgments containing the outputs of 14 models on SAMSum.

DACSA: A large-scale Dataset for Automatic summarization of Catalan and Spanish newspaper Articles
Encarnación Segarra Soriano, Vicent Ahuir, Lluís-F. Hurtado and José Ángel González 08:45-09:00 (Columbia C)
The application of supervised methods to automatic summarization requires the availability of adequate corpora consisting of a set of document-summary pairs. As in most Natural Language Processing tasks, the great majority of available datasets for summarization are in English, making it difficult to develop automatic summarization models for other languages. Although Spanish is gradually forming part of some recent summarization corpora, it is not the same for minority languages such as Catalan. In this work, we describe the construction of a corpus of Catalan and Spanish newspapers, the Dataset for Automatic summarization of Catalan and Spanish newspaper Articles (DACSA) corpus. It is a high-quality large-scale corpus that can be used to train summarization models for Catalan and Spanish. We have carried out an analysis of the corpus, both in terms of the style of the summaries and the difficulty of the summarization task. In particular, we have used a set of well-known metrics in the summarization field in order to characterize the corpus. Additionally, for benchmarking purposes, we have evaluated the performances of some extractive and abstractive summarization systems on the DACSA corpus.

Information Extraction 2

08:00-09:00 (Columbia D)

Robust Self-Augmentation for Named Entity Recognition with Meta Reweighting
Linzhi Wu, Pengjun Xie, Jie Zhou, Meishan Zhang, Ma Chunping, Guangwei Xu and Min Zhang 08:00-08:15 (Columbia D)
Self-augmentation has received increasing research interest recently to improve named entity recognition (NER) performance in low-resource scenarios. Token substitution and mixup are two feasible heterogeneous self-augmentation techniques for NER that can achieve effective performance with certain specialized efforts. Noticeably, self-augmentation may introduce potentially noisy augmented data. Prior research has mainly resorted to heuristic rule-based constraints to reduce the noise for specific self-augmentation methods individually. In this paper, we revisit these two typical self-augmentation methods for NER, and propose a unified meta-reweighting strategy for them to achieve a natural integration. Our method is easily extensible, imposing little effort on a specific self-augmentation method. Experiments on different Chinese and English NER benchmarks show that our token substitution and mixup method, as well as their integration, can achieve effective performance improvement. Based on the meta-reweighting mechanism, we can enhance the advantages of the self-augmentation techniques without much extra effort.

GMN: Generative Multi-modal Network for Practical Document Information Extraction
Haoyu Cao, Jiefeng Ma, Antai Guo, Yiqing Hu, Hao Liu, Deqiang Jiang, Yinsong Liu and Bo Ren 08:15-08:30 (Columbia D)
Document Information Extraction (DIE) has attracted increasing attention due to its various advanced applications in the real world. Although recent literature has already achieved competitive results, these approaches usually fail when dealing with complex documents with noisy OCR results or mutative layouts. This paper proposes Generative Multi-modal Network (GMN) for real-world scenarios to address these problems, which is a robust multi-modal generation method without predefined label categories. With the carefully designed spatial encoder and modal-aware mask module, GMN can deal with complex documents that are hard to serialized into sequential order. Moreover, GMN tolerates errors in OCR results and requires no character-level annotation, which is vital because fine-grained annotation of numerous documents is laborious and even requires annotators with specialized domain knowledge. Extensive experiments show that GMN achieves new state-of-the-art performance on several public DIE datasets and surpasses other methods by a large margin, especially in realistic scenes.

DocEE: A Large-Scale and Fine-grained Benchmark for Document-level Event Extraction
MeiHan Tong, Bin Xu, Shuai Wang, Meihuan Han, Yixin Cao, Jiangqi Zhu, Siyu Chen, Lei Hou and Juanzi Li 08:30-08:45 (Columbia D)
Event extraction aims to identify an event and then extract the arguments participating in the event. Despite the great success in sentence-level event extraction, events are more naturally presented in the form of documents, with event arguments scattered in multiple sentences. However, a major barrier to promote document-level event extraction has been the lack of large-scale and practical training and evaluation datasets. In this paper, we present DocEE, a new document-level event extraction dataset including 27,000+ events, 180,000+ arguments. We highlight three features: large-scale manual annotations, fine-grained argument types and application-oriented settings. Experiments show that there is still a big gap between state-of-the-art models and human beings (41% Vs 85% in F1 score), indicating that DocEE is an open issue. DocEE is now available at https://github.com/tongmeihan1995/DocEE.git.

HiURE: Hierarchical Exemplar Contrastive Learning for Unsupervised Relation Extraction
Shuliang Liu, Xuming Hu, Chenwei Zhang, Shu'ang Li, Lijie Wen and Philip S. Yu 08:45-09:00 (Columbia D)
Unsupervised relation extraction aims to extract the relationship between entities from natural language sentences without prior information on relational scope or distribution. Existing works either utilize self-supervised schemes to refine relational feature signals by iteratively leveraging adaptive clustering and classification that provoke gradual drift problems, or adopt instance-wise contrastive learning which unreasonably pushes apart those sentence pairs that are semantically similar. To overcome these defects, we propose a novel contrastive learning framework named HiURE, which has the capability to derive hierarchical signals from relational feature space using cross hierarchy attention and effectively optimize relation representation of sentences under exemplar-wise contrastive learning. Experimental results on two public datasets demonstrate the advanced effectiveness and robustness of HiURE on unsupervised relation extraction when compared with state-of-

the-art models.

Machine Translation 2

08:00-09:00 (Elwha A)

Neighbors Are Not Strangers: Improving Non-Autoregressive Translation under Low-Frequency Lexical Constraints
Chun Zeng, Jiangjie Chen, Tianyi Zhuang, Rui Xu, Hao Yang, Qin Ying, Shimin Tao and Yanghua Xiao 08:00-08:15 (Elwha A)
Lexically constrained neural machine translation (NMT) draws much industrial attention for its practical usage in specific domains. However, current autoregressive approaches suffer from high latency. In this paper, we focus on non-autoregressive translation (NAT) for this problem for its efficiency advantage. We identify that current constrained NAT models, which are based on iterative editing, do not handle low-frequency constraints well. To this end, we propose a plug-in algorithm for this line of work, i.e., Aligned Constrained Training (ACT), which alleviates this problem by familiarizing the model with the source-side context of the constraints. Experiments on the general and domain datasets show that our model improves over the backbone constrained NAT model in constraint preservation and translation quality, especially for rare constraints.

Nearest Neighbor Knowledge Distillation for Neural Machine Translation
Zhixian Yang, Renliang Sun and Xiaojun Wan 08:15-08:30 (Elwha A)
k-nearest-neighbor machine translation (kNN-MT), proposed by Khandelwal et al. (2021), has achieved many state-of-the-art results in machine translation tasks. Although effective, kNN-MT requires conducting kNN searches through the large datastore for each decoding step during inference, prohibitively increasing the decoding cost and thus leading to the difficulty for the deployment in real-world applications. In this paper, we propose to move the time-consuming kNN search forward to the preprocessing phase, and then introduce k Nearest Neighbor Knowledge Distillation (kNN-KD) that trains the base NMT model to directly learn the knowledge of kNN. Distilling knowledge retrieved by kNN can encourage the NMT model to take more reasonable target tokens into consideration, thus addressing the overcorrection problem. Extensive experimental results show that, the proposed method achieves consistent improvement over the state-of-the-art baselines including kNN-MT, while maintaining the same training and decoding speed as the standard NMT model.

Cross-modal Contrastive Learning for Speech Translation
Rong Ye, Mingxuan Wang and Lei Li 08:30-08:45 (Elwha A)
How can we learn unified representations for spoken utterances and their written text? Learning similar representations for semantically similar speech and text is important for speech translation. To this end, we propose ConST, a cross-modal contrastive learning method for end-to-end speech-to-text translation. We evaluate ConST and a variety of previous baselines on a popular benchmark MuST-C. Experiments show that the proposed ConST consistently outperforms the previous methods, and achieves an average BLEU of 29.4. The analysis further verifies that ConST indeed closes the representation gap of different modalities — its learned representation improves the accuracy of cross-modal speech-text retrieval from 4% to 88%. Code and models are available at https://github.com/ReneeYe/ConST.

One Reference Is Not Enough: Diverse Distillation with Reference Selection for Non-Autoregressive Translation
Chenze Shao, Xuanfu Wu and Yang Feng 08:45-09:00 (Elwha A)
Non-autoregressive neural machine translation (NAT) suffers from the multi-modality problem: the source sentence may have multiple correct translations, but the loss function is calculated only according to the reference sentence. Sequence-level knowledge distillation makes the target more deterministic by replacing the target with the output from an autoregressive model. However, the multi-modality problem in the distilled dataset is still nonnegligible. Furthermore, learning from a specific teacher limits the upper bound of the model capability, restricting the potential of NAT models. In this paper, we argue that one reference is not enough and propose diverse distillation with reference selection (DDRS) for NAT. Specifically, we first propose a method called SeedDiv for diverse machine translation, which enables us to generate a dataset containing multiple high-quality reference translations for each source sentence. During the training, we compare the NAT output with all references and select the one that best fits the NAT output to train the model. Experiments on widely-used machine translation benchmarks demonstrate the effectiveness of DDRS, which achieves 29.82 BLEU with only one decoding pass on WMT14 En-De, improving the state-of-the-art performance for NAT by over 1 BLEU.

Dialogue and Interactive Systems 2

08:00-09:00 (Elwha B)

Less is More: Learning to Refine Dialogue History for Personalized Dialogue Generation
Hanxun Zhong, Zhicheng Dou, Yutao Zhu, Hongjin Qian and Ji-Rong Wen 08:00-08:15 (Elwha B)
Personalized dialogue systems explore the problem of generating responses that are consistent with the user's personality, which has raised much attention in recent years. Existing personalized dialogue systems have tried to extract user profiles from dialogue history to guide personalized response generation. Since the dialogue history is usually long and noisy, most existing methods truncate the dialogue history to model the user's personality. Such methods can generate some personalized responses, but a large part of dialogue history is wasted, leading to sub-optimal performance of personalized response generation. In this work, we propose to refine the user dialogue history on a large scale, based on which we can handle more dialogue history and obtain more abundant and accurate persona information. Specifically, we design an MSP model which consists of three personal information refiners and a personalized response generator. With these multi-level refiners, we can sparsely extract the most valuable information (tokens) from the dialogue history and leverage other similar users' data to enhance personalization. Experimental results on two real-world datasets demonstrate the superiority of our model in generating more informative and personalized responses.

Diversifying Neural Dialogue Generation via Negative Distillation
Yiwei Li, Shaoxiong Feng, Bin Sun and Kan Li 08:15-08:30 (Elwha B)
Generative dialogue models suffer badly from the generic response problem, limiting their applications to a few toy scenarios. Recently, an interesting approach, namely negative training, has been proposed to alleviate this problem by reminding the model not to generate high-frequency responses during training. However, its performance is hindered by two issues, ignoring low-frequency but generic responses and bringing low-frequency but meaningless responses. In this paper, we propose a novel negative training paradigm, called negative distillation, to keep the model away from the undesirable generic responses while avoiding the above problems. First, we introduce a negative teacher

model that can produce query-wise generic responses, and then the student model is required to maximize the distance with multi-level negative knowledge. Empirical results show that our method outperforms previous negative training methods significantly.

Learning as Conversation: Dialogue Systems Reinforced for Information Acquisition
Pengshan Cai, Hui Wan, Fei Liu, Mo Yu, Hong yu and Sachindra Joshi 08:30-08:45 (Elwha B)
We propose novel AI-empowered chat bots for learning as conversation where a user does not read a passage but gains information and knowledge through conversation with a teacher bot. Our information acquisition-oriented dialogue system employs a novel adaptation of reinforced self-play so that the system can be transferred to various domains without in-domain dialogue data, and can carry out conversations both informative and attentive to users.

Enhancing Knowledge Selection for Grounded Dialogues via Document Semantic Graphs
Sha Li, Mahdi Namazifar, Di Jin, Mohit Bansal, Heng Ji, Yang Liu and Dilek Hakkani-Tur 08:45-09:00 (Elwha B)
Providing conversation models with background knowledge has been shown to make open-domain dialogues more informative and engaging. Existing models treat knowledge selection as a sentence ranking or classification problem where each sentence is handled individually, ignoring the internal semantic connection between sentences. In this work, we propose to automatically convert the background knowledge documents into document semantic graphs and then perform knowledge selection over such graphs. Our document semantic graphs preserve sentence-level information through the use of sentence nodes and provide concept connections between sentences. We apply multi-task learning to perform sentence-level knowledge selection and concept-level knowledge selection, showing that it improves sentence-level selection. Our experiments show that our semantic graph-based knowledge selection improves over sentence selection baselines for both the knowledge selection task and the end-to-end response generation task on HollE and improves generalization on unseen topics in WoW.

Machine Learning for NLP 2

08:00-09:00 (Quinault)

LaMemo: Language Modeling with Look-Ahead Memory
Haozhe Ji, Rongsheng Zhang, Zhenyu Yang, Zhipeng Hu and Minlie Huang 08:00-08:15 (Quinault)
Although Transformers with fully connected self-attentions are powerful to model long-term dependencies, they are struggling to scale to long texts with thousands of words in language modeling. One of the solutions is to equip the model with a recurrence memory. However, existing approaches directly reuse hidden states from the previous segment that encodes contexts in a uni-directional way. As a result, this prohibits the memory to dynamically interact with the current context that provides up-to-date information for token prediction. To remedy this issue, we propose Look-Ahead Memory (LaMemo) that enhances the recurrence memory by incrementally attending to the right-side tokens and interpolating with the old memory states to maintain long-term information in the history. LaMemo embraces bi-directional attention and segment recurrence with an additional computation overhead only linearly proportional to the memory length. Experiments on widely used language modeling benchmarks demonstrate its superiority over the baselines equipped with different types of memory mechanisms.

[TACL] **Formal Language Recognition by Hard Attention Transformers: Perspectives from Circuit Complexity**
Yiding Hao, Dana Angluin and Robert Evan Frank 08:15-08:30 (Quinault)
This paper analyzes three formal models of Transformer encoders that differ in the form of their self-attention mechanism: unique hard attention (UHAT), generalized unique hard attention (GUHAT), which generalizes UHAT, and averaging hard attention (AHAT). We show that Transformers in UHAT and GUHAT can recognize only formal languages recognizable by families of Boolean circuits of constant depth and polynomial size–that is, families of circuits in the complexity class AC 0 . By the classic results of Furst et al. (1984), this upper bound subsumes Hahn (2020)'s results that GUHAT cannot recognize the DYCK languages or the PARITY language. In contrast, the languages MAJORITY and DYCK-1, non-AC 0 languages, are recognizable by AHAT networks, implying that AHAT can recognize languages that UHAT and GUHAT cannot.

Symbolic Knowledge Distillation: from General Language Models to Commonsense Models
*Peter West, Chandra Bhagavatula, Jack Hessel, Jena D. Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck and Yejin Choi*08:30-08:45 (Quinault)
The common practice for training commonsense models has gone from–human–to–corpus–to–machine: humans author commonsense knowledge graphs in order to train commonsense models. In this work, we investigate an alternative, from–machine–to–corpus–to–machine: general language models author these commonsense knowledge graphs to train commonsense models. Our study leads to a new framework, Symbolic Knowledge Distillation. As with prior art in Knowledge Distillation (Hinton et al. 2015), our approach uses larger models to teach smaller models. A key difference is that we distill knowledge symbolically–as text–in addition to the neural model. We distill only one aspect–the commonsense of a general language model teacher, allowing the student to be a different type, a commonsense model. Altogether, we show that careful prompt engineering and a separately trained critic model allow us to selectively distill high-quality causal commonsense from GPT-3, a general language model. Empirical results demonstrate that, for the first time, a human-authored commonsense knowledge graph is surpassed by our automatically distilled variant in all three criteria: quantity, quality, and diversity. In addition, it results in a neural commonsense model that surpasses the teacher model's commonsense capabilities despite its 100x smaller size. We apply this to the ATOMIC resource, and will share our new symbolic knowledge graph and commonsense models.

WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models
Benjamin Minixhofer, Fabian Paischer and Navid Rekabsaz 08:45-09:00 (Quinault)
Large pretrained language models (LMs) have become the central building block of many NLP applications. Training these models requires ever more computational resources and most of the existing models are trained on English text only. It is exceedingly expensive to train these models in other languages. To alleviate this problem, we introduce a novel method – called WECHSEL – to efficiently and effectively transfer pretrained LMs to new languages. WECHSEL can be applied to any model which uses subword-based tokenization and learns an embedding for each subword. The tokenizer of the source model (in English) is replaced with a tokenizer in the target language and token embeddings are initialized such that they are semantically similar to the English tokens by utilizing multilingual static word embeddings covering English and the target language. We use WECHSEL to transfer the English RoBERTa and GPT-2 models to four languages (French, German, Chinese and Swahili). We also study the benefits of our method on very low-resource languages. WECHSEL improves over proposed methods for cross-lingual parameter transfer and outperforms models of comparable size trained from scratch with up to 64x less training effort. Our method makes training large language models for new languages more accessible and less damaging to the environment. We make our code and models publicly available.

Virtual Poster Q&A Session 3

08:00-09:00 (702 Clearwater)

Political Ideology and Polarization: A Multi-dimensional Approach
Barea Sinno, Bernardo Oviedo, Katherine Atwell, Malihe Alikhani and Junyi Jessy Li 08:00-09:00 (702 Clearwater)
Analyzing ideology and polarization is of critical importance in advancing our grasp of modern politics. Recent research has made great strides towards understanding the ideological bias (i.e., stance) of news media along the left-right spectrum. In this work, we instead take a novel and more nuanced approach for the study of ideology based on its left or right positions on the issue being discussed. Aligned with the theoretical accounts in political science, we treat ideology as a multi-dimensional construct, and introduce the first diachronic dataset of news articles whose ideological positions are annotated by trained political scientists and linguists at the paragraph level. We showcase that, by controlling for the author's stance, our method allows for the quantitative and temporal measurement and analysis of polarization as a multidimensional ideological distance. We further present baseline models for ideology prediction, outlining a challenging task distinct from stance detection.

Combining Humor and Sarcasm for Improving Political Parody Detection
Xiao Ao, Danae Sanchez Villegas, Daniel Preotiuc-Pietro and Nikolaos Aletras 08:00-09:00 (702 Clearwater)
Parody is a figurative device used for mimicking entities for comedic or critical purposes. Parody is intentionally humorous and often involves sarcasm. This paper explores jointly modelling these figurative tropes with the goal of improving performance of political parody detection in tweets. To this end, we present a multi-encoder model that combines three parallel encoders to enrich parody-specific representations with humor and sarcasm information. Experiments on a publicly available data set of political parody tweets demonstrate that our approach outperforms previous state-of-the-art methods.

Counterfactually Augmented Data and Unintended Bias: The Case of Sexism and Hate Speech Detection
Indira Sen, Mattia Samory, Claudia Wagner and Isabelle Augenstein 08:00-09:00 (702 Clearwater)
Counterfactually Augmented Data (CAD) aims to improve out-of-domain generalizability, an indicator of model robustness. The improvement is credited to promoting core features of the construct over spurious artifacts that happen to correlate with it. Yet, over-relying on core features may lead to unintended model bias. Especially, construct-driven CAD—perturbations of core features—may induce models to ignore the context in which core features are used. Here, we test models for sexism and hate speech detection on challenging data: non-hate and non-sexist usage of identity and gendered terms. On these hard cases, models trained on CAD, especially construct-driven CAD, show higher false positive rates than models trained on the original, unperturbed data. Using a diverse set of CAD—construct-driven and construct-agnostic—reduces such unintended bias.

Conceptualizing Treatment Leakage in Text-based Causal Inference
Adel Daoud, Connor Thomas Jerzak and Richard Johansson 08:00-09:00 (702 Clearwater)
Causal inference methods that control for text-based confounders are becoming increasingly important in the social sciences and other disciplines where text is readily available. However, these methods rely on a critical assumption that there is no treatment leakage: that is, the text only contains information about the confounder and no information about treatment assignment. When this assumption does not hold, methods that control for text to adjust for confounders face the problem of post-treatment (collider) bias. However, the assumption that there is no treatment leakage may be unrealistic in real-world situations involving text, as human language is rich and flexible. Language appearing in a public policy document or health records may refer to the future and the past simultaneously, and thereby reveal information about the treatment assignment.

In this article, we define the treatment-leakage problem, and discuss the identification as well as the estimation challenges it raises. Second, we delineate the conditions under which leakage can be addressed by removing the treatment-related signal from the text in a pre-processing step we define as text distillation. Lastly, using simulation, we show how treatment leakage introduces a bias in estimates of the average treatment effect (ATE) and how text distillation can mitigate this bias.

DISARM: Detecting the Victims Targeted by Harmful Memes
Shivam Sharma, Md Shad Akhtar, Preslav Nakov and Tanmoy Chakraborty 08:00-09:00 (702 Clearwater)
Internet memes have emerged as an increasingly popular means of communication on the web. Although memes are typically intended to elicit humour, they have been increasingly used to spread hatred, trolling, and cyberbullying, as well as to target specific individuals, communities, or society on political, socio-cultural, and psychological grounds. While previous work has focused on detecting harmful, hateful, and offensive memes in general, identifying whom these memes attack (i.e., the 'victims') remains a challenging and underexplored area. We attempt to address this problem in this paper. To this end, we create a dataset in which we annotate each meme with its victim(s) such as the name of the targeted person(s), organization(s), and community(ies). We then propose DISARM (Detecting vIctimS targeted by hARmful Memes), a framework that uses named-entity recognition and person identification to detect all entities a meme is referring to, and then, incorporates a novel contextualized multimodal deep neural network to classify whether the meme intends to harm these entities. We perform several systematic experiments on three different test sets, corresponding to entities that are (i) all seen while training, (ii) not seen as a harmful target while training, and (iii) not seen at all while training. The evaluation shows that DISARM significantly outperforms 10 unimodal and multimodal systems. Finally, we demonstrate that DISARM is interpretable and comparatively more generalizable and that it can reduce the relative error rate of harmful target identification by up to 9 % absolute over multimodal baseline systems.

Analyzing the Intensity of Complaints on Social Media
Ming Fang, Shi Zong, Jing Li, Xinyu Dai, Shujian Huang and Jiajun Chen 08:00-09:00 (702 Clearwater)
Complaining is a speech act that expresses a negative inconsistency between reality and human's expectations. While prior studies mostly focus on identifying the existence or the type of complaints, in this work, we present the first study in computational linguistics of measuring the intensity of complaints from text. Analyzing complaints from such perspective is particularly useful, as complaints of certain degrees may cause severe consequences for companies or organizations. We first collect 3,103 posts about complaints in education domain from Weibo, a popular Chinese social media platform. These posts are then annotated with complaints intensity scores using Best-Worst Scaling (BWS) method. We show that complaints intensity can be accurately estimated by computational models with best mean square error achieving 0.11. Furthermore, we conduct a comprehensive linguistic analysis around complaints, including the connections between complaints and sentiment, and a cross-lingual comparison for complaints expressions used by Chinese and English speakers. We finally show that our complaints intensity scores can be incorporated for better estimating the popularity of posts on social media.

CRUSH: Contextually Regularized and User anchored Self-supervised Hate speech Detection
Souvic Chakraborty, Parag Dutta, Sumegh Roychowdhury and Animesh Mukherjee 08:00-09:00 (702 Clearwater)
The last decade has witnessed a surge in the interaction of people through social networking platforms. While there are several positive aspects of these social platforms, their proliferation has led them to become the breeding ground for cyber-bullying and hate speech. Recent advances

in NLP have often been used to mitigate the spread of such hateful content. Since the task of hate speech detection is usually applicable in the context of social networks, we introduce CRUSH, a framework for hate speech detection using User Anchored self-supervision and contextual regularization. Our proposed approach secures 1-12% improvement in test set metrics over best performing previous approaches on two types of tasks and multiple popular English language social networking datasets.

Triggerless Backdoor Attack for NLP Tasks with Clean Labels
Leilei Gan, Jiwei Li, Tianwei Zhang, Xiaoya Li, Yuxian Meng, Fei Wu, Yi Yang, Shangwei Guo and Chun Fan 08:00-09:00 (702 Clearwater)
Backdoor attacks pose a new threat to NLP models. A standard strategy to construct poisoned data in backdoor attacks is to insert triggers (e.g., rare words) into selected sentences and alter the original label to a target label. This strategy comes with a severe flaw of being easily detected from both the trigger and the label perspectives: the trigger injected, which is usually a rare word, leads to an abnormal natural language expression, and thus can be easily detected by a defense model; the changed target label leads the example to be mistakenly labeled, and thus can be easily detected by manual inspections. To deal with this issue, in this paper, we propose a new strategy to perform textual backdoor attack which does not require an external trigger and the poisoned samples are correctly labeled. The core idea of the proposed strategy is to construct clean-labeled examples, whose labels are correct but can lead to test label changes when fused with the training set. To generate poisoned clean-labeled examples, we propose a sentence generation model based on the genetic algorithm to cater to the non-differentiable characteristic of text data. Extensive experiments demonstrate that the proposed attacking strategy is not only effective, but more importantly, hard to defend due to its triggerless and clean-labeled nature. Our work marks the first step towards developing triggerless attacking strategies in NLP.

An Information-Theoretic Approach and Dataset for Probing Gender Stereotypes in Multilingual Masked Language Models
Victor Steinborn, Philipp Dufter, Haris Jabbar and Hinrich Schuetze 08:00-09:00 (702 Clearwater)
Bias research in NLP is a rapidly growing and developing field. Similar to CrowS-Pairs (Nangia et al., 2020), we assess gender bias in masked-language models (MLMs) by studying pairs of sentences with gender swapped person references. Most bias research focuses on and often is specific to English. Using a novel methodology for creating sentence pairs that is applicable across languages, we create, based on CrowS-Pairs, a multilingual dataset for English, Finnish, German, Indonesian and Thai. Additionally, we propose S_JSD, a new bias measure based on Jensen–Shannon divergence, which we argue retains more information from the model output probabilities than other previously proposed bias measures for MLMs. Using multilingual MLMs, we find that S_JSD diagnoses the same systematic biased behavior for non-English that previous studies have found for monolingual English pre-trained MLMs. S_JSD outperforms the CrowS-Pairs measure, which struggles to find such biases for smaller non-English datasets.

What Makes a Good and Useful Summary? Incorporating Users in Automatic Summarization Research
Maartje Ter Hoeve, Julia Kiseleva and Maarten de Rijke 08:00-09:00 (702 Clearwater)
Automatic text summarization has enjoyed great progress over the years and is used in numerous applications, impacting the lives of many. Despite this development, there is little research that meaningfully investigates how the current research focus in automatic summarization aligns with users' needs. To bridge this gap, we propose a survey methodology that can be used to investigate the needs of users of automatically generated summaries. Importantly, these needs are dependent on the target group. Hence, we design our survey in such a way that it can be easily adjusted to investigate different user groups. In this work we focus on university students, who make extensive use of summaries during their studies. We find that the current research directions of the automatic summarization community do not fully align with students' needs. Motivated by our findings, we present ways to mitigate this mismatch in future research on automatic summarization: we propose research directions that impact the design, the development and the evaluation of automatically generated summaries.

Quiz Design Task: Helping Teachers Create Quizzes with Automated Question Generation
Philippe Laban, Chien-Sheng Wu, Lidiya Murakhovs'ka, Wenhao Liu and Caiming Xiong 08:00-09:00 (702 Clearwater)
Question generation (QGen) models are often evaluated with standardized NLG metrics that are based on n-gram overlap. In this paper, we measure whether these metric improvements translate to gains in a practical setting, focusing on the use case of helping teachers automate the generation of reading comprehension quizzes. In our study, teachers building a quiz receive question suggestions, which they can either accept or refuse with a reason. Even though we find that recent progress in QGen leads to a significant increase in question acceptance rates, there is still large room for improvement, with the best model having only 68.4% of its questions accepted by the ten teachers who participated in our study. We then leverage the annotations we collected to analyze standard NLG metrics and find that model performance has reached projected upper-bounds, suggesting new automatic metrics are needed to guide QGen research forward.

Label Definitions Improve Semantic Role Labeling
Li Zhang, Ishan Jindal and Yunyao Li 08:00-09:00 (702 Clearwater)
Argument classification is at the core of Semantic Role Labeling. Given a sentence and the predicate, a semantic role label is assigned to each argument of the predicate. While semantic roles come with meaningful definitions, existing work has treated them as symbolic. Learning symbolic labels usually requires ample training data, which is frequently unavailable due to the cost of annotation. We instead propose to retrieve and leverage the definitions of these labels from the annotation guidelines. For example, the verb predicate "work" has arguments defined as "worker", "job", "employer", etc. Our model achieves state-of-the-art performance on the CoNLL09 dataset injected with label definitions given the predicate senses. The performance improvement is even more pronounced in low-resource settings when training data is scarce.

FOAM: A Follower-aware Speaker Model For Vision-and-Language Navigation
Zi-Yi Dou and Nanyun Peng 08:00-09:00 (702 Clearwater)
The speaker-follower models have proven to be effective in vision-and-language navigation, where a speaker model is used to synthesize new instructions to augment the training data for a follower navigation model. However, in previous work, the speaker model is follower-agnostic and fails to take the state of the follower into consideration. In this paper, we present FOAM, a FOllower-Aware speaker Model that is constantly updated given the follower feedback, so that the generated instructions can be more suitable to the current learning state of the follower. Specifically, we optimize the speaker using a bi-level optimization framework and obtain its training signals by evaluating the follower on labeled data. Experimental results on the Room-to-Room and Room-across-Room datasets demonstrate that our methods can outperform strong baseline models across settings. Analyses also reveal that our generated instructions are of higher quality than the baselines.

KD-VLP: Improving End-to-End Vision-and-Language Pretraining with Object Knowledge Distillation
Yongfei Liu, Chenfei Wu, Shao-Yen Tseng, Vasudev Lal, Xuming He and Nan Duan 08:00-09:00 (702 Clearwater)
Self-supervised vision-and-language pretraining (VLP) aims to learn transferable multi-modal representations from large-scale image-text data and to achieve strong performances on a broad scope of vision-language tasks after finetuning. Previous mainstream VLP approaches typically adopt a two-step strategy relying on external object detectors to encode images in a multi-modal Transformer framework, which suffer from restrictive object concept space, limited image context and inefficient computation. In this paper, we propose an object-aware end-to-end VLP framework, which directly feeds image grid features from CNNs into the Transformer and learns the multi-modal representations jointly. More importantly, we propose to perform object knowledge distillation to facilitate learning cross-modal alignment at different

semantic levels. To achieve that, we design two novel pretext tasks by taking object features and their semantic labels from external detectors as supervision: 1.) Object-guided masked vision modeling task focuses on enforcing object-aware representation learning in the multi-modal Transformer; 2.) Phrase-region alignment task aims to improve cross-modal alignment by utilizing the similarities between noun phrases and object labels in the linguistic space. Extensive experiments on a wide range of vision-language tasks demonstrate the efficacy of our proposed framework, and we achieve competitive or superior performances over the existing pretraining strategies.

Cross-Lingual Cross-Modal Consolidation for Effective Multilingual Video Corpus Moment Retrieval
Jiaheng Liu, Tan Yu, Hanyu Peng, Mingming Sun and Ping Li 08:00-09:00 (702 Clearwater)
Existing multilingual video corpus moment retrieval (mVCMR) methods are mainly based on a two-stream structure. The visual stream utilizes the visual content in the video to estimate the query-visual similarity, and the subtitle stream exploits the query-subtitle similarity. The final query-video similarity ensembles similarities from two streams. In our work, we pro- pose a simple and effective strategy termed as Cross-lingual Cross-modal Consolidation (C3) to improve mVCMR accuracy. We adopt the ensemble similarity as the teacher to guide the training of each stream, leading to a more powerful ensemble similarity. Meanwhile, we use the teacher for a specific language to guide the student for another language to exploit the complementary knowledge across languages. Ex- tensive experiments on mTVR dataset demonstrate the effectiveness of our C3 method.

Beyond Emotion: A Multi-Modal Dataset for Human Desire Understanding
Ao Jia, Yu He, Yazhou Zhang, Sagar Uprety, Dawei Song and Christina Lioma 08:00-09:00 (702 Clearwater)
Desire is a strong wish to do or have something, which involves not only a linguistic expression, but also underlying cognitive phenomena driving human feelings. As the most primitive and basic human instinct, conscious desire is often accompanied by a range of emotional responses. As a strikingly understudied task, it is difficult for machines to model and understand desire due to the unavailability of benchmarking datasets with desire and emotion labels. To bridge this gap, we present MSED, the first multi-modal and multi-task sentiment, emotion and desire dataset, which contains 9,190 text-image pairs, with English text. Each multi-modal sample is annotated with six desires, three sentiments and six emotions. We also propose the state-of-the-art baselines to evaluate the potential of MSED and show the importance of multi-task and multi-modal clues for desire understanding. We hope this study provides a benchmark for human desire analysis. MSED will be publicly available for research.

Are All the Datasets in Benchmark Necessary? A Pilot Study of Dataset Evaluation for Text Classification
Yang Xiao, Jinlan Fu, See-Kiong Ng and Pengfei Liu 08:00-09:00 (702 Clearwater)
In this paper, we ask the research question of whether all the datasets in the benchmark are necessary. We approach this by first characterizing the distinguishability of datasets when comparing different systems. Experiments on 9 datasets and 36 systems show that several existing benchmark datasets contribute little to discriminating top-scoring systems, while those less used datasets exhibit impressive discriminative power. We further, taking the text classification task as a case study, investigate the possibility of predicting dataset discrimination based on its properties (e.g., average sentence length). Our preliminary experiments promisingly show that given a sufficient number of training experimental records, a meaningful predictor can be learned to estimate dataset discrimination over unseen datasets. We released all datasets with features explored in this work on DataLab.

MultiNERD: A Multilingual, Multi-Genre and Fine-Grained Dataset for Named Entity Recognition (and Disambiguation)
Simone Tedeschi and Roberto Navigli 08:00-09:00 (702 Clearwater)
Named Entity Recognition (NER) is the task of identifying named entities in texts and classifying them through specific semantic categories, a process which is crucial for a wide range of NLP applications. Current datasets for NER focus mainly on coarse-grained entity types, tend to consider a single textual genre and to cover a narrow set of languages, thus limiting the general applicability of NER systems. In this work, we design a new methodology for automatically producing NER annotations, and address the aforementioned limitations by introducing a novel dataset that covers 10 languages, 15 NER categories and 2 textual genres. We also introduce a manually-annotated test set, and extensively evaluate the quality of our novel dataset on both this new test set and standard benchmarks for NER. In addition, in our dataset, we include: i) disambiguation information to enable the development of multilingual entity linking systems, and ii) image URLs to encourage the creation of multimodal systems. We release our dataset at https://github.com/Babelscape/multinerd.

ID10M: Idiom Identification in 10 Languages
Simone Tedeschi, Federico Martelli and Roberto Navigli 08:00-09:00 (702 Clearwater)
Idioms are phrases which present a figurative meaning that cannot be (completely) derived by looking at the meaning of their individual components. Identifying and understanding idioms in context is a crucial goal and a key challenge in a wide range of Natural Language Understanding tasks. Although efforts have been undertaken in this direction, the automatic identification and understanding of idioms is still a largely under-investigated area, especially when operating in a multilingual scenario. In this paper, we address such limitations and put forward several new contributions: we propose a novel multilingual Transformer-based system for the identification of idioms; we produce a high-quality automatically-created training dataset in 10 languages, along with a novel manually-curated evaluation benchmark; finally, we carry out a thorough performance analysis and release our evaluation suite at https://github.com/Babelscape/ID10M.

User-Driven Research of Medical Note Generation Software
Tom Knoll, Francesco Moramarco, Alex Papadopoulos Korfiatis, Rachel Young, Claudia Ruffini, Mark Perera, Christian Perstl, Ehud Reiter, Anya Belz and Aleksandar Savkov 08:00-09:00 (702 Clearwater)
A growing body of work uses Natural Language Processing (NLP) methods to automatically generate medical notes from audio recordings of doctor-patient consultations. However, there are very few studies on how such systems could be used in clinical practice, how clinicians would adjust to using them, or how system design should be influenced by such considerations. In this paper, we present three rounds of user studies, carried out in the context of developing a medical note generation system. We present, analyse and discuss the participating clinicians' impressions and views of how the system ought to be adapted to be of value to them. Next, we describe a three-week test run of the system in a live telehealth clinical practice. Major findings include (i) the emergence of five different note-taking behaviours; (ii) the importance of the system generating notes in real time during the consultation; and (iii) the identification of a number of clinical use cases that could prove challenging for automatic note generation systems.

Re2G: Retrieve, Rerank, Generate
Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Naik, Pengshan Cai and Alfio Gliozzo 08:00-09:00 (702 Clearwater)
As demonstrated by GPT-3 and T5, transformers grow in capability as parameter spaces become larger and larger. However, for tasks that require a large amount of knowledge, non-parametric memory allows models to grow dramatically with a sub-linear increase in computational cost and GPU memory requirements. Recent models such as RAG and REALM have introduced retrieval into conditional generation. These models incorporate neural initial retrieval from a corpus of passages. We build on this line of research, proposing Re2G, which combines both neural initial retrieval and reranking into a BART-based sequence-to-sequence generation. Our reranking approach also permits merging retrieval results from sources with incomparable scores, enabling an ensemble of BM25 and neural initial retrieval. To train our system end-

to-end, we introduce a novel variation of knowledge distillation to train the initial retrieval, reranker and generation using only ground truth on the target sequence output. We find large gains in four diverse tasks: zero-shot slot filling, question answering, fact checking and dialog, with relative gains of 9% to 34% over the previous state-of-the-art on the KILT leaderboard. We make our code available as open source.

Seeing the wood for the trees: a contrastive regularization method for the low-resource Knowledge Base Question Answering
jpliu@wtu.edu.cn jpliu@wtu.edu.cn, Shijie Mei, Xinrong Hu, Xun Yao, Jack Yang and Yi Guo 08:00-09:00 (702 Clearwater)
Given a context knowledge base (KB) and a corresponding question, the Knowledge Base Question Answering task aims to retrieve correct answer entities from this KB. Despite sophisticated retrieval algorithms, the impact of the low-resource (incomplete) KB is not fully exploited, where contributing components (i.e. key entities and/or relations) may be absent for question answering. To effectively address this problem, we propose a contrastive regularization based method, which is motivated by the `learn-by-analogy` capability from human readers. Specifically, the proposed work includes two major modules: the knowledge extension and sMoCo module. The former aims at exploiting the latent knowledge from the context KB and generating auxiliary information in the form of question-answer pairs. The later module utilizes those additional pairs and applies the contrastive regularization to learn informative representations, that making `hard` positive pairs attracted and `hard` negative pairs separated. Empirically, we achieved the state-of-the-art performance on the WebQuestionsSP dataset and the effectiveness of proposed modules is also evaluated.

To Answer or Not To Answer? Improving Machine Reading Comprehension Model with Span-based Contrastive Learning
Yunjie Ji, Liangyu Chen, Chenxiao Dou, Baochang Ma and Xiangang Li 08:00-09:00 (702 Clearwater)
Machine Reading Comprehension with Unanswerable Questions is a difficult NLP task, challenged by the questions which can not be answered from passages. It is observed that subtle literal changes often make an answerable question unanswerable, however, most MRC models fail to recognize such changes. To address this problem, in this paper, we propose a span-based method of Contrastive Learning (spanCL) which explicitly contrast answerable questions with their answerable and unanswerable counterparts at the answer span level. With spanCL, MRC models are forced to perceive crucial semantic changes from slight literal differences. Experiments on SQuAD 2.0 dataset show that spanCL can improve baselines significantly, yielding 0.86~2.14 absolute EM improvements. Additional experiments also show that spanCL is an effective way to utilize generated questions.

All Information is Valuable: Question Matching over Full Information Transmission Network
Le Qi, Yu Zhang, Qingyu Yin, Guidong Zheng, Wen Junjie, Jinlong Li and Ting Liu 08:00-09:00 (702 Clearwater)
Question matching is the task of identifying whether two questions have the same intent. For better reasoning the relationship between questions, existing studies adopt multiple interaction modules and perform multi-round reasoning via deep neural networks. In this process, there are two kinds of critical information that are commonly employed: the representation information of original questions and the interactive information between pairs of questions. However, previous studies tend to transmit only one kind of information, while failing to utilize both kinds of information simultaneously. To address this problem, in this paper, we propose a Full Information Transmission Network (FITN) that can transmit both representation and interactive information together in a simultaneous fashion. More specifically, we employ a novel memory-based attention for keeping and transmitting the interactive information through a global interaction matrix. Besides, we apply an original-average mixed connection method to effectively transmit the representation information between different reasoning rounds, which helps to preserve the original representation features of questions along with the historical hidden features. Experiments on two standard benchmarks demonstrate that our approach outperforms strong baseline models.

Great Truths are Always Simple **: A Rather Simple Knowledge Encoder for Enhancing the Commonsense Reasoning Capacity of Pre-Trained Models**
Jinhao Jiang, Kun Zhou, Ji-Rong Wen and Xin Zhao 08:00-09:00 (702 Clearwater)
Commonsense reasoning in natural language is a desired ability of artificial intelligent systems. For solving complex commonsense reasoning tasks, a typical solution is to enhance pre-trained language models (PTMs) with a knowledge-aware graph neural network (GNN) encoder that models a commonsense knowledge graph (CSKG). Despite the effectiveness, these approaches are built on heavy architectures, and can't clearly explain how external knowledge resources improve the reasoning capacity of PTMs. Considering this issue, we conduct a deep empirical analysis, and find that it is indeed *relation features* from CSKGs (but not *node features*) that mainly contribute to the performance improvement of PTMs. Based on this finding, we design a simple MLP-based knowledge encoder that utilizes statistical relation paths as features. Extensive experiments conducted on five benchmarks demonstrate the effectiveness of our approach, which also largely reduces the parameters for encoding CSKGs. Our codes and data are publicly available at `https://github.com/RUCAIBox/SAFE`.

Capturing Conversational Interaction for Question Answering via Global History Reasoning
Jin Qian, Bowei Zou, Mengxing Dong, Xiao Li, AiTi Aw and Yu Hong 08:00-09:00 (702 Clearwater)
Conversational Question Answering (ConvQA) is required to answer the current question, conditioned on the observable paragraph-level context and conversation history. Previous works have intensively studied history-dependent reasoning. They perceive and absorb topic-related information of prior utterances in the interactive encoding stage. It yielded significant improvement compared to history-independent reasoning. This paper further strengthens the ConvQA encoder by establishing long-distance dependency among global utterances in multi-turn conversation. We use multi-layer transformers to resolve long-distance relationships, which potentially contribute to the reweighting of attentive information in historical utterances. Experiments on QuAC show that our method obtains a substantial improvement (1%), yielding the F1 score of 73.7%. All source codes are available at https://github.com/jaytsien/GHR.

Continual Machine Reading Comprehension via Uncertainty-aware Fixed Memory and Adversarial Domain Adaptation
Zhijing Wu, Hua Xu, Jingliang Fang and Kai Gao 08:00-09:00 (702 Clearwater)
Continual Machine Reading Comprehension aims to incrementally learn from a continuous data stream across time without access the previous seen data, which is crucial for the development of real-world MRC systems. However, it is a great challenge to learn a new domain incrementally without catastrophically forgetting previous knowledge. In this paper, MA-MRC, a continual MRC model with uncertainty-aware fixed Memory and Adversarial domain adaptation, is proposed. In MA-MRC, a fixed size memory stores a small number of samples in previous domain data along with an uncertainty-aware updating strategy when new domain data arrives. For incremental learning, MA-MRC not only keeps a stable understanding by learning both memory and new domain data, but also makes full use of the domain adaptation relationship between them by adversarial learning strategy. The experimental results show that MA-MRC is superior to strong baselines and has a substantial incremental learning ability without catastrophically forgetting under two different continual MRC settings.

CIAug: Equipping Interpolative Augmentation with Curriculum Learning
Ramit Sawhney, Ritesh Singh Soun, Shrey Pandit, Megh Thakkar, Sarvagya Malaviya and Yuval Pinter 08:00-09:00 (702 Clearwater)
Interpolative data augmentation has proven to be effective for NLP tasks. Despite its merits, the sample selection process in mixup is random, which might make it difficult for the model to generalize better and converge faster. We propose CIAug, a novel curriculum-based learning method that builds upon mixup. It leverages the relative position of samples in hyperbolic embedding space as a complexity measure to gradually mix up increasingly difficult and diverse samples along training. CIAug achieves state-of-the-art results over existing interpolative augmentation methods on 10 benchmark datasets across 4 languages in text classification and named-entity recognition tasks. It also con-

verges and achieves benchmark F1 scores 3 times faster. We empirically analyze the various components of CIAug, and evaluate its robustness against adversarial attacks.

EPiDA: An Easy Plug-in Data Augmentation Framework for High Performance Text Classification
Minyi Zhao, Lu Zhang, Yi Xu, Jiandong Ding, Jihong Guan and Shuigeng Zhou 08:00-09:00 (702 Clearwater)
Recent works have empirically shown the effectiveness of data augmentation (DA) in NLP tasks, especially for those suffering from data scarcity. Intuitively, given the size of generated data, their diversity and quality are crucial to the performance of targeted tasks. However, to the best of our knowledge, most existing methods consider only either the diversity or the quality of augmented data, thus cannot fully mine the potential of DA for NLP. In this paper, we present an easy and plug-in data augmentation framework EPiDA to support effective text classification. EPiDA employs two mechanisms: relative entropy maximization (REM) and conditional entropy minimization (CEM) to control data generation, where REM is designed to enhance the diversity of augmented data while CEM is exploited to ensure their semantic consistency. EPiDA can support efficient and continuous data generation for effective classifier training. Extensive experiments show that EPiDA outperforms existing SOTA methods in most cases, though not using any agent networks or pre-trained generation networks, and it works well with various DA algorithms and classification models.

Enhancing Self-Attention with Knowledge-Assisted Attention Maps
Jiangang Bai, Yujing Wang, Hong Sun, Ruonan Wu, Tianmeng Yang, Pengfei Tang, Defu Cao, Mingliang Zhang1, Yunhai Tong, Yaming Yang, Jing Bai, Ruofei Zhang, Hao Sun and Wei Shen 08:00-09:00 (702 Clearwater)
Large-scale pre-trained language models have attracted extensive attentions in the research community and shown promising results on various tasks of natural language processing. However, the attention maps, which record the attention scores between tokens in self-attention mechanism, are sometimes ineffective as they are learned implicitly without the guidance of explicit semantic knowledge. Thus, we aim to infuse explicit external knowledge into pre-trained language models to further boost their performance. Existing works of knowledge infusion largely depend on multi-task learning frameworks, which are inefficient and require large-scale re-training when new knowledge is considered. In this paper, we propose a novel and generic solution, KAM-BERT, which directly incorporates knowledge-generated attention maps into the self-attention mechanism. It requires only a few extra parameters and supports efficient fine-tuning once new knowledge is added. KAM-BERT achieves consistent improvements on various academic datasets for natural language understanding. It also outperforms other state-of-the-art methods which conduct knowledge infusion into transformer-based architectures. Moreover, we apply our model to an industry-scale ad relevance application and show its advantages in the real-world scenario.

Natural Language Inference with Self-Attention for Veracity Assessment of Pandemic Claims
Miguel Arana-Catania, Elena Kochkina, Arkaitz Zubiaga, Maria Liakata, Robert Procter and Yulan He 08:00-09:00 (702 Clearwater)
We present a comprehensive work on automated veracity assessment from dataset creation to developing novel methods based on Natural Language Inference (NLI), focusing on misinformation related to the COVID-19 pandemic. We first describe the construction of the novel PANACEA dataset consisting of heterogeneous claims on COVID-19 and their respective information sources. The dataset construction includes work on retrieval techniques and similarity measurements to ensure a unique set of claims. We then propose novel techniques for automated veracity assessment based on Natural Language Inference including graph convolutional networks and attention based approaches. We have carried out experiments on evidence retrieval and veracity assessment on the dataset using the proposed techniques and found them competitive with SOTA methods, and provided a detailed discussion.

ValCAT: Variable-Length Contextualized Adversarial Transformations Using Encoder-Decoder Language Model
Chuyun Deng, Mingxuan Liu, Yue Qin, Jia Zhang, Hai-Xin Duan and Donghong Sun 08:00-09:00 (702 Clearwater)
Adversarial texts help explore vulnerabilities in language models, improve model robustness, and explain their working mechanisms. However, existing word-level attack methods trap in a one-to-one attack pattern, i.e., only a single word can be modified in one transformation round, and they ignore the interactions between several consecutive words. In this paper, we propose ValCAT, a black-box attack framework that misleads the language model by applying variable-length contextualized transformations to the original text. Compared to word-level methods, ValCAT expands the basic units of perturbation from single words to spans composed of multiple consecutive words, enhancing the perturbation capability. Experiments show that our method outperforms state-of-the-art methods in terms of attack success rate, perplexity, and semantic similarity on several classification tasks and inference tasks. The comprehensive human evaluation demonstrates that ValCAT has a significant advantage in ensuring the fluency of the adversarial examples and achieves better semantic consistency. We release the code at https://github.com/linerxliner/ValCAT.

DynamicTOC: Persona-based Table of Contents for Consumption of Long Documents
Himanshu Maheshwari, Nethraa Sivakumar, Shelly Jain, Tanvi Karandikar, Vinay Aggarwal, Navita Goyal and Sumit Shekhar 08:00-09:00 (702 Clearwater)
Long documents like contracts, financial documents, etc., are often tedious to read through. Linearly consuming (via scrolling or navigation through default table of content) these documents is time-consuming and challenging. These documents are also authored to be consumed by varied entities (referred to as persona in the paper) interested in only certain parts of the document. In this work, we describe DynamicToC, a dynamic table of content-based navigator, to aid in the task of non-linear, persona-based document consumption. DynamicToC highlights sections of interest in the document as per the aspects relevant to different personas. DynamicToC is augmented with short questions to assist the users in understanding underlying content. This uses a novel deep-reinforcement learning technique to generate questions on these persona-clustered paragraphs. Human and automatic evaluations suggest the efficacy of both end-to-end pipeline and different components of DynamicToC.

Non-Autoregressive Chinese ASR Error Correction with Phonological Training
Zheng Fang, Ruiqing Zhang, Zhongjun He, Hua Wu and Yanan Cao 08:00-09:00 (702 Clearwater)
Automatic Speech Recognition (ASR) is an efficient and widely used input method that transcribes speech signals into text. As the errors introduced by ASR systems will impair the performance of downstream tasks, we introduce a post-processing error correction method, PhVEC, to correct errors in text space. For the errors in ASR result, existing works mainly focus on fixed-length corrections, modifying each wrong token to a correct one (one-to-one correction), but rarely consider the variable-length correction (one-to-many or many-to-one correction). In this paper, we propose an efficient non-autoregressive (NAR) method for Chinese ASR error correction for both cases. Instead of conventionally predicting the sentence length in NAR methods, we propose a novel approach that uses phonological tokens to extend the source sentence for variable-length correction, enabling our model to generate phonetically similar corrections. Experimental results on datasets of different domains show that our method achieves significant improvement in word error rate reduction and speeds up the inference by 6.2 times compared with the autoregressive model.

Measuring and Improving Compositional Generalization in Text-to-SQL via Component Alignment
Yujian Gan, Xinyun Chen, Qiuping Huang and Matthew Purver 08:00-09:00 (702 Clearwater)
In text-to-SQL tasks — as in much of NLP — *compositional generalization* is a major challenge: neural networks struggle with compositional generalization where training and test distributions differ. However, most recent attempts to improve this are based on word-level synthetic

data or specific dataset splits to generate compositional biases. In this work, we propose a clause-level compositional example generation method. We first split the sentences in the Spider text-to-SQL dataset into sub-sentences, annotating each sub-sentence with its corresponding SQL clause, resulting in a new dataset Spider-SS. We then construct a further dataset, Spider-CG, by composing Spider-SS sub-sentences in different combinations, to test the ability of models to generalize compositionally. Experiments show that existing models suffer significant performance degradation when evaluated on Spider-CG, even though every sub-sentence is seen during training. To deal with this problem, we modify a number of state-of-the-art models to train on the segmented data of Spider-SS, and we show that this method improves the generalization performance.

CODE-MVP: Learning to Represent Source Code from Multiple Views with Contrastive Pre-Training
Xin Wang, Yasheng Wang, Yao Wan, Jiawei Wang, Pingyi Zhou, Li Li, Hao Wu and Jin Liu 08:00-09:00 (702 Clearwater)
Recent years have witnessed increasing interest in code representation learning, which aims to represent the semantics of source code into distributed vectors. Currently, various works have been proposed to represent the complex semantics of source code from different views, including plain text, Abstract Syntax Tree (AST), and several kinds of code graphs (e.g., Control/Data Flow Graph). However, most of them only consider a single view of source code independently, ignoring the correspondences among different views. In this paper, we propose to integrate different views with the natural-language description of source code into a unified framework with Multi-View contrastive Pre-training, and name our model as CODE-MVP. Specifically, we first extract multiple code views using compiler tools, and learn the complementary information among them under a contrastive learning framework. Inspired by the type checking in compilation, we also design a fine-grained type inference objective in the pre-training. Experiments on three downstream tasks over five datasets demonstrate the superiority of CODE-MVP when compared with several state-of-the-art baselines. For example, we achieve 2.4/2.3/1.1 gain in terms of MRR/MAP/Accuracy metrics on natural language code retrieval, code similarity, and code defect detection tasks, respectively.

Unbiased Math Word Problems Benchmark for Mitigating Solving Bias
ZhiCheng Yang, Jinghui Qin, Jiaqi Chen and Xiaodan Liang 08:00-09:00 (702 Clearwater)
In this paper, we revisit the solving bias when evaluating models on current Math Word Problem (MWP) benchmarks. However, current solvers exist solving bias which consists of data bias and learning bias due to biased dataset and improper training strategy. Our experiments verify MWP solvers are easy to be biased by the biased training datasets which do not cover diverse questions for each problem narrative of all MWPs, thus a solver can only learn shallow heuristics rather than deep semantics for understanding problems. Besides, an MWP can be naturally solved by multiple equivalent equations while current datasets take only one of the equivalent equations as ground truth, forcing the model to match the labeled ground truth and ignoring other equivalent equations. Here, we first introduce a novel MWP dataset named UnbiasedMWP which is constructed by varying the grounded expressions in our collected data and annotating them with corresponding multiple new questions manually. Then, to further mitigate learning bias, we propose a Dynamic Target Selection (DTS) Strategy to dynamically select more suitable target expressions according to the longest prefix match between the current model output and candidate equivalent equations which are obtained by applying commutative law during training. The results show that our UnbiasedMWP has significantly fewer biases than its original data and other datasets, posing a promising benchmark for fairly evaluating the solvers' reasoning skills rather than matching nearest neighbors. And the solvers trained with our DTS achieve higher accuracies on multiple MWP benchmarks. The source code is available at https://github.com/yangzhch6/UnbiasedMWP.

Pathway2Text: Dataset and Method for Biomedical Pathway Description Generation
Junwei Yang, Zequn Liu, Ming Zhang and Sheng Wang 08:00-09:00 (702 Clearwater)
Biomedical pathways have been extensively used to characterize the mechanism of complex diseases. One essential step in biomedical pathway analysis is to curate the description of a pathway based on its graph structure and node features. Neural text generation could be a plausible technique to circumvent the tedious manual curation. In this paper, we propose a new dataset Pathway2Text, which contains 2,367 pairs of biomedical pathways and textual descriptions. All pathway graphs are experimentally derived or manually curated. All textual descriptions are written by domain experts. We form this problem as a Graph2Text task and propose a novel graph-based text generation approach kNN-Graph2Text, which explicitly exploited descriptions of similar graphs to generate new descriptions. We observed substantial improvement of our method on both Graph2Text and the reverse task of Text2Graph. We further illustrated how our dataset can be used as a novel benchmark for biomedical named entity recognition. Collectively, we envision our method will become an important benchmark for evaluating Graph2Text methods and advance biomedical research for complex diseases.

D2GCLF: Document-to-Graph Classifier for Legal Document Classification
Qiqi Wang, Kaiqi Zhao, Robert Amor, Benjamin Liu and Ruofan Wang 08:00-09:00 (702 Clearwater)
Legal document classification is an essential task in law intelligence to automate the labor-intensive law case filing process. Unlike traditional document classification problems, legal documents should be classified by reasons and facts instead of topics. We propose a Document-to-Graph Classifier (D2GCLF), which extracts facts as relations between key participants in the law case and represents a legal document with four relation graphs. Each graph is responsible for capturing different relations between the litigation participants. We further develop a graph attention network on top of the four relation graphs to classify the legal documents. Experiments on a real-world legal document dataset show that D2GCLF outperforms the state-of-the-art methods in terms of accuracy.

Explore More Guidance: A Task-aware Instruction Network for Sign Language Translation Enhanced with Data Augmentation
Yong Cao, Wei Li, Xianzhi Li, Min Chen, Guangyong Chen, Long Hu, Zhengdao Li and Kai Hwang 08:00-09:00 (702 Clearwater)
Sign language recognition and translation first uses a recognition module to generate glosses from sign language videos and then employs a translation module to translate glosses into spoken sentences. Most existing works focus on the recognition step, while paying less attention to sign language translation. In this work, we propose a task-aware instruction network, namely TIN-SLT, for sign language translation, by introducing the isntruction module and the learning-based feature fuse strategy into a Transformer network. In this way, the pre-trained model's language ability can be well explored and utilized to further boost the translation performance. Moreover, by exploring the representation space of sign language glosses and target spoken language, we propose a multi-level data augmentation scheme to adjust the data distribution of the training set. We conduct extensive experiments on two challenging benchmark datasets, PHOENIX-2014-T and ASLG-PC12, on which our method outperforms former best solutions by 1.65 and 1.42 in terms of BLEU-4. Our code and trained networks will be available upon the publication of this work.

Query2Particles: Knowledge Graph Reasoning with Particle Embeddings
Jiaxin Bai, Zihao Wang, Hongming Zhang and Yangqiu Song 08:00-09:00 (702 Clearwater)
Answering complex logical queries on incomplete knowledge graphs (KGs) with missing edges is a fundamental and important task for knowledge graph reasoning. The query embedding method is proposed to answer these queries by jointly encoding queries and entities to the same embedding space. Then the answer entities are selected according to the similarities between the entity embeddings and the query embedding. As the answers to a complex query are obtained from a combination of logical operations over sub-queries, the embeddings of the answer entities may not always follow a uni-modal distribution in the embedding space. Thus, it is challenging to simultaneously retrieve a set of diverse answers from the embedding space using a single and concentrated query representation such as a vector or a hyper-rectangle. To better cope with queries with diversified answers, we propose Query2Particles (Q2P), a complex KG query answering method. Q2P encodes

each query into multiple vectors, named particle embeddings. By doing so, the candidate answers can be retrieved from different areas over the embedding space using the maximal similarities between the entity embeddings and any of the particle embeddings. Meanwhile, the corresponding neural logic operations are defined to support its reasoning over arbitrary first-order logic queries. The experiments show that Query2Particles achieves state-of-the-art performance on the complex query answering tasks on FB15k, FB15K-237, and NELL knowledge graphs.

Towards Job-Transition-Tag Graph for a Better Job Title Representation Learning
Jun Zhu and Celine Hudelot 08:00-09:00 (702 Clearwater)
Works on learning job title representation are mainly based on *Job-Transition Graph*, built from the working history of talents. However, since these records are usually messy, this graph is very sparse, which affects the quality of the learned representation and hinders further analysis. To address this specific issue, we propose to enrich the graph with additional nodes that improve the quality of job title representation. Specifically, we construct *Job-Transition-Tag Graph*, a heterogeneous graph containing two types of nodes, i.e., job titles and tags (i.e., words related to job responsibilities or functionalities). Along this line, we reformulate job title representation learning as the task of learning node embedding on the *Job-Transition-Tag Graph*. Experiments on two datasets show the interest of our approach.

Aspect Is Not You Need: No-aspect Differential Sentiment Framework for Aspect-based Sentiment Analysis
Jiahao Cao, Rui Liu, Huailiang Peng, Lei Jiang and Xu Bai 08:00-09:00 (702 Clearwater)
Aspect-based sentiment analysis (ABSA) is a fine-grained sentiment classification task. Most recent efforts adopt pre-trained model to classify the sentences with aspects. However, the aspect sentiment bias from pre-trained model brings some noise to the ABSA task. Besides, traditional methods using cross-entropy loss are hard to find the potential associations between sentiment polarities. In this work, we analyze the ABSA task from a novel cognition perspective: humans can often judge the sentiment of an aspect even if they do not know what the aspect is. Moreover, it is easier to distinguish positive and negative sentiments than others for human beings because positive and negative are two opposite sentiments. To this end, we propose a no-aspect differential sentiment (NADS) framework for the ABSA task. We first design a no-aspect template by replacing the aspect with a special unbiased character to eliminate the sentiment bias and obtain a stronger representation. To better get the benefits from the template, we adopt contrastive learning between the no-aspect template and the original sentence. Then we propose a differential sentiment loss instead of the cross-entropy loss to better classify the sentiments by distinguishing the different distances between sentiments. Our proposed model is a general framework and can be combined with almost all traditional ABSA methods. Experiments on SemEval 2014 show that our framework is still able to predict the sentiment of the aspect even we don't konw what the aspect is. Moreover, our NADS framework boosts three typical ABSA methods and achieves state-of-the-art performance.

Generative Cross-Domain Data Augmentation for Aspect and Opinion Co-Extraction
Junjie Li, Jianfei Yu and Rui Xia 08:00-09:00 (702 Clearwater)
As a fundamental task in opinion mining, aspect and opinion co-extraction aims to identify the aspect terms and opinion terms in reviews. However, due to the lack of fine-grained annotated resources, it is hard to train a robust model for many domains. To alleviate this issue, unsupervised domain adaptation is proposed to transfer knowledge from a labeled source domain to an unlabeled target domain. In this paper, we propose a new Generative Cross-Domain Data Augmentation framework for unsupervised domain adaptation. The proposed framework is aimed to generate target-domain data with fine-grained annotation by exploiting the labeled data in the source domain. Specifically, we remove the domain-specific segments in a source-domain labeled sentence, and then use this as input to a pre-trained sequence-to-sequence model BART to simultaneously generate a target-domain sentence and predict the corresponding label for each word. Experimental results on three datasets demonstrate that our approach is more effective than previous domain adaptation methods.

A Dual-Channel Framework for Sarcasm Recognition by Detecting Sentiment Conflict
Yiyi Liu, Yequan Wang, Aixin Sun, Xuying Meng, Jing Li and Jiafeng Guo 08:00-09:00 (702 Clearwater)
Sarcasm employs ambivalence, where one says something positive but actually means negative, and vice versa. The essence of sarcasm, which is also a sufficient and necessary condition, is the conflict between literal and implied sentiments expressed in one sentence. However, it is difficult to recognize such sentiment conflict because the sentiments are mixed or even implicit. As a result, the recognition of sophisticated and obscure sentiment brings in a great challenge to sarcasm detection. In this paper, we propose a Dual-Channel Framework by modeling both literal and implied sentiments separately. Based on this dual-channel framework, we design the Dual-Channel Network (DC-Net) to recognize sentiment conflict. Experiments on political debates (i.e. IAC-V1 and IAC-V2) and Twitter datasets show that our proposed DC-Net achieves state-of-the-art performance on sarcasm recognition. Our code is released to support research.

CLMLF:A Contrastive Learning and Multi-Layer Fusion Method for Multimodal Sentiment Detection
Zhen Li, Bing Xu, Conghui Zhu and Tiejun Zhao 08:00-09:00 (702 Clearwater)
Compared with unimodal data, multimodal data can provide more features to help the model analyze the sentiment of data. Previous research works rarely consider token-level feature fusion, and few works explore learning the common features related to sentiment in multimodal data to help the model fuse multimodal features. In this paper, we propose a Contrastive Learning and Multi-Layer Fusion (CLMLF) method for multimodal sentiment detection. Specifically, we first encode text and image to obtain hidden representations, and then use a multi-layer fusion module to align and fuse the token-level features of text and image. In addition to the sentiment analysis task, we also designed two contrastive learning tasks, label based contrastive learning and data based contrastive learning tasks, which will help the model learn common features related to sentiment in multimodal data. Extensive experiments conducted on three publicly available multimodal datasets demonstrate the effectiveness of our approach for multimodal sentiment detection compared with existing methods. The codes are available for use at https: //github.com/Link-Li/CLMLF

Leaner and Faster: Two-Stage Model Compression for Lightweight Text-Image Retrieval
Siyu Ren and Kenny Q. Zhu 08:00-09:00 (702 Clearwater)
Current text-image approaches (e.g., CLIP) typically adopt dual-encoder architecture using pre-trained vision-language representation. However, these models still pose non-trivial memory requirements and substantial incremental indexing time, which makes them less practical on mobile devices. In this paper, we present an effective two-stage framework to compress large pre-trained dual-encoder for lightweight text-image retrieval. The resulting model is smaller (39
faster (1.6x/2.9x for processing image/text respectively), yet performs on par with or better than the original full model on Flickr30K and MSCOCO benchmarks. We also open-source an accompanying realistic mobile image search application.

Exact Paired-Permutation Testing for Structured Test Statistics
Ran Zmigrod, Tim Vieira and Ryan Cotterell 08:00-09:00 (702 Clearwater)
Significance testing—especially the paired-permutation test—has played a vital role in developing NLP systems to provide confidence that the difference in performance between two systems (i.e., the test statistic) is not due to luck. However, practitioners rely on Monte Carlo approximation to perform this test due to a lack of a suitable exact algorithm. In this paper, we provide an efficient exact algorithm for the paired-permutation test for a family of structured test statistics. Our algorithm runs in $\mathcal{O}(GN(\log GN)(\log N))$ time where N is the dataset size and G is the range of the test statistic. We found that our exact algorithm was 10x faster than the Monte Carlo approximation

with 20000 samples on a common dataset

Efficient Learning of Multiple NLP Tasks via Collective Weight Factorization on BERT
Christos Charalampos Papadopoulos, Yannis Panagakis, Manolis Koubarakis and Mihalis Nicolaou 08:00-09:00 (702 Clearwater)
The Transformer architecture continues to show remarkable performance gains in many Natural Language Processing tasks. However, obtaining such state-of-the-art performance in different tasks requires fine-tuning the same model separately for each task. Clearly, such an approach is demanding in terms of both memory requirements and computing power. In this paper, aiming to improve training efficiency across multiple tasks, we propose to collectively factorize the weighs of the multi-head attention module of a pre-trained Transformer. We test our proposed method on finetuning multiple natural language understanding tasks by employing BERT-Large as an instantiation of the Transformer and the GLUE as the evaluation benchmark. Experimental results show that our method requires training and storing only 1% of the initial model parameters for each task and matches or improves the original fine-tuned model's performance for each task while effectively decreasing the parameter requirements by two orders of magnitude. Furthermore, compared to well-known adapter-based alternatives on the GLUE benchmark, our method consistently reaches the same levels of performance while requiring approximately four times fewer total and trainable parameters per task.

RAIL-KD: RAndom Intermediate Layer Mapping for Knowledge Distillation
Md Akmal Haidar, Nithin Anchuri, Mehdi Rezagholizadeh, Abbas Ghaddar, Philippe Langlais and Pascal Poupart 08:00-09:00 (702 Clearwater)
Intermediate layer knowledge distillation (KD) can improve the standard KD technique (which only targets the output of teacher and student models) especially over large pre-trained language models. However, intermediate layer distillation suffers from excessive computational burdens and engineering efforts required for setting up a proper layer mapping. To address these problems, we propose a RAndom Intermediate Layer Knowledge Distillation (RAIL-KD) approach in which, intermediate layers from the teacher model are selected randomly to be distilled into the intermediate layers of the student model. This randomized selection enforces that all teacher layers are taken into account in the training process, while reducing the computational cost of intermediate layer distillation. Also, we show that it acts as a regularizer for improving the generalizability of the student model. We perform extensive experiments on GLUE tasks as well as on out-of-domain test sets. We show that our proposed RAIL-KD approach outperforms other state-of-the-art intermediate layer KD methods considerably in both performance and training-time.

PCEE-BERT: Accelerating BERT Inference via Patient and Confident Early Exiting
Zhen Zhang, Wei Zhu, Jinfan Zhang, Peng Wang, Rize Jin and Tae-Sun Chung 08:00-09:00 (702 Clearwater)
BERT and other pretrained language models (PLMs) are ubiquitous in modern NLP. Even though PLMs are the state-of-the-art (SOTA) models for almost every NLP task, the significant latency during inference prohibits wider industrial usage. In this work, we propose Patient and Confident Early Exiting BERT (PCEE-BERT), an off-the-shelf sample-dependent early exiting method that can work with different PLMs and can also work along with popular model compression methods. With a multi-exit BERT as the backbone model, PCEE-BERT will make the early exiting decision if enough numbers (patience parameter) of consecutive intermediate layers are confident about their predictions. The entropy value measures the confidence level of an intermediate layer's prediction. Experiments on the GLUE benchmark demonstrate that our method outperforms previous SOTA early exiting methods. Ablation studies show that: (a) our method performs consistently well on other PLMs, such as ALBERT and TinyBERT; (b) PCEE-BERT can achieve different speed-up ratios by adjusting the patience parameter and the confidence threshold. The code for PCEE-BERT can be found at `https://github.com/michael-wzhu/PCEE-BERT`.

Cross-Domain Detection of GPT-2-Generated Technical Text
Juan Diego Rodriguez, Todd Hay, David Gros, Zain Shamsi and Ravi Srinivasan 08:00-09:00 (702 Clearwater)
Machine-generated text presents a potential threat not only to the public sphere, but also to the scientific enterprise, whereby genuine research is undermined by convincing, synthetic text. In this paper we examine the problem of detecting GPT-2-generated technical research text. We first consider the realistic scenario where the defender does not have full information about the adversary's text generation pipeline, but is able to label small amounts of in-domain genuine and synthetic text in order to adapt to the target distribution. Even in the extreme scenario of adapting a physics-domain detector to a biomedical detector, we find that only a few hundred labels are sufficient for good performance. Finally, we show that paragraph-level detectors can be used to detect the tampering of full-length documents under a variety of threat models.

Revisiting Generative Commonsense Reasoning: A Pre-Ordering Approach
Chao Zhao, Faeze Brahman, Tenghao Huang and Snigdha Chaturvedi 08:00-09:00 (702 Clearwater)
Pre-trained models (PTMs) have lead to great improvements in natural language generation (NLG). However, it is still unclear how much commonsense knowledge they possess. With the goal of evaluating commonsense knowledge of NLG models, recent work has proposed the problem of generative commonsense reasoning, e.g., to compose a logical sentence given a set of unordered concepts. Existing approaches to this problem hypothesize that PTMs lack sufficient parametric knowledge for this task, which can be overcome by introducing external knowledge or task-specific pre-training objectives. Different from this trend, we argue that PTM's inherent ability for generative commonsense reasoning is underestimated due to the order-agnostic property of its input. In particular, we hypothesize that the order of the input concepts can affect the PTM's ability to utilize its commonsense knowledge. To this end, we propose a pre-ordering approach to elaborately manipulate the order of the given concepts before generation. Experiments show that our approach can outperform the more sophisticated models that have access to a lot of external data and resources.

Learning from Bootstrapping and Stepwise Reinforcement Reward: A Semi-Supervised Framework for Text Style Transfer
Zhengyuan Liu and Nancy F. Chen 08:00-09:00 (702 Clearwater)
Text style transfer is an important task in controllable language generation. Supervised approaches have pushed performance improvement on style-oriented rewriting such as formality conversion. However, challenges remain due to the scarcity of large-scale parallel data in many domains. While unsupervised approaches do not rely on annotated sentence pairs for each style, they are often plagued with instability issues such as mode collapse or quality degradation. To take advantage of both supervised and unsupervised paradigms and tackle the challenges, in this work, we propose a semi-supervised framework for text style transfer. First, the learning process is bootstrapped with supervision guided by automatically constructed pseudo-parallel pairs using lexical and semantic-based methods. Then the model learns from unlabeled data via reinforcement rewards. Specifically, we propose to improve the sequence-to-sequence policy gradient via stepwise reward optimization, providing fine-grained learning signals and stabilizing the reinforced learning process. Experimental results show that the proposed approach achieves state-of-the-art performance on multiple datasets, and produces effective generation with as minimal as 10% of training data.

Unsupervised Domain Adaptation for Question Generation with DomainData Selection and Self-training
Peide Zhu and Claudia Hauff 08:00-09:00 (702 Clearwater)
Question generation (QG) approaches based on large neural models require (i) large-scale and (ii) high-quality training data. These two requirements pose difficulties for specific application domains where training data is expensive and difficult to obtain. The trained QG models' effectiveness can degrade significantly when they are applied on a different domain due to domain shift. In this paper, we explore an *unsupervised domain adaptation* approach to combat the lack of training data and domain shift issue with domain data selection and self-training. We

first present a novel answer-aware strategy for domain data selection to select data with the most similarity to a new domain. The selected data are then used as pseudo-in-domain data to retrain the QG model. We then present generation confidence guided self-training with two generation confidence modeling methods (i) generated questions' perplexity and (ii) the fluency score. We test our approaches on three large public datasets with different domain similarities, using a transformer-based pre-trained QG model. The results show that our proposed approaches outperform the baselines, and show the viability of unsupervised domain adaptation with answer-aware data selection and self-training on the QG task.

Learning Structural Information for Syntax-Controlled Paraphrase Generation
Erguang Yang, Chenglin Bai, Deyi Xiong, Yujie Zhang, Yao Meng, Jinan Xu and Yufeng Chen 08:00-09:00 (702 Clearwater)
Syntax-controlled paraphrase generation aims to produce paraphrase conform to given syntactic patterns. To address this task, recent works have started to use parse trees (or syntactic templates) to guide generation. A constituency parse tree contains abundant structural information, such as parent-child relation, sibling relation, and the alignment relation between words and nodes. Previous works have only utilized parent-child and alignment relations, which may affect the generation quality. To address this limitation, we propose a Structural Information-augmented Syntax-Controlled Paraphrasing (SI-SCP) model. Particularly, we design a syntax encoder based on tree-transformer to capture parent-child and sibling relations. To model the alignment relation between words and nodes, we propose an attention regularization objective, which makes the decoder accurately select corresponding syntax nodes to guide the generation of words. Experiments show that SI-SCP achieves state-of-the-art performances in terms of semantic and syntactic quality on two popular benchmark datasets. Additionally, we propose a Syntactic Template Retriever (STR) to retrieve compatible syntactic structures. We validate that STR is capable of retrieving compatible syntactic structures. We further demonstrate the effectiveness of SI-SCP to generate diverse paraphrases with retrieved syntactic structures.

Quantifying Language Variation Acoustically with Few Resources
Martijn Bartelds and Martijn Wieling 08:00-09:00 (702 Clearwater)
Deep acoustic models represent linguistic information based on massive amounts of data. Unfortunately, for regional languages and dialects such resources are mostly not available. However, deep acoustic models might have learned linguistic information that transfers to low-resource languages. In this study, we evaluate whether this is the case through the task of distinguishing low-resource (Dutch) regional varieties. By extracting embeddings from the hidden layers of various wav2vec 2.0 models (including new models which are pre-trained and/or fine-tuned on Dutch) and using dynamic time warping, we compute pairwise pronunciation differences averaged over 10 words for over 100 individual dialects from four (regional) languages. We then cluster the resulting difference matrix in four groups and compare these to a gold standard, and a partitioning on the basis of comparing phonetic transcriptions. Our results show that acoustic models outperform the (traditional) transcription-based approach without requiring phonetic transcriptions, with the best performance achieved by the multilingual XLSR-53 model fine-tuned on Dutch. On the basis of only six seconds of speech, the resulting clustering closely matches the gold standard.

FAtNet: Cost-Effective Approach Towards Mitigating the Linguistic Bias in Speaker Verification Systems
Divya V Sharma and Arun Balaji Buduru 08:00-09:00 (702 Clearwater)
Linguistic bias in Deep Neural Network (DNN) based Natural Language Processing (NLP) systems is a critical problem that needs attention. The problem further intensifies in the case of security systems, such as speaker verification, where fairness is essential. Speaker verification systems are intelligent systems that determine if two speech recordings belong to the same speaker. Such human-oriented security systems should be usable by diverse people speaking varied languages. Thus, a speaker verification system trained on speech in one language should generalize when tested for other languages. However, DNN-based models are often language-dependent. Previous works explore domain adaptation to fine-tune the pre-trained model for out-of-domain languages. Fine-tuning the model individually for each existing language is expensive. Hence, it limits the usability of the system. This paper proposes the cost-effective idea of integrating a lightweight embedding with existing speaker verification systems to mitigate linguistic bias without adaptation. This work is motivated by the theoretical hypothesis that attentive-frames could help generate language-agnostic embeddings. For scientific validation of this hypothesis, we propose two frame-attentive networks and investigate the effect of their integration with baselines for twelve languages. Empirical results suggest that frame-attentive embedding can cost-effectively reduce linguistic bias and enhance the usability of baselines.

Penn-Helsinki Parsed Corpus of Early Modern English: First Parsing Results and Analysis
Seth Kulick, Neville Ryant and Beatrice Santorini 08:00-09:00 (702 Clearwater)
The Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME), a 1.7-million-word treebank that is an important resource for research in syntactic change, has several properties that present potential challenges for NLP technologies. We describe these key features of PPCEME that make it challenging for parsing, including a larger and more varied set of function tags than in the Penn Treebank, and present results for this corpus using a modified version of the Berkeley Neural Parser and the approach to function tag recovery of Gabbard et al. (2006). While this approach to function tag recovery gives reasonable results, it is in some ways inappropriate for span-based parsers. We also present further evidence of the importance of in-domain pretraining for contextualized word representations. The resulting parser will be used to parse Early English Books Online, a 1.5 billion word corpus whose utility for the study of syntactic change will be greatly increased with the addition of accurate parse trees.

SHARP: Search-Based Adversarial Attack for Structured Prediction
Liwen Zhang, Zixia Jia, Wenjuan Han, Zilong Zheng and Kewei Tu 08:00-09:00 (702 Clearwater)
Adversarial attack of structured prediction models faces various challenges such as the difficulty of perturbing discrete words, the sentence quality issue, and the sensitivity of outputs to small perturbations. In this work, we introduce SHARP, a new attack method that formulates the black-box adversarial attack as a search-based optimization problem with a specially designed objective function considering sentence fluency, meaning preservation and attacking effectiveness. Additionally, three different searching strategies are analyzed and compared, i.e., Beam Search, Metropolis-Hastings Sampling, and Hybrid Search. We demonstrate the effectiveness of our attacking strategies on two challenging structured prediction tasks: Pos-tagging and dependency parsing. Through automatic and human evaluations, we show that our method performs a more potent attack compared with pioneer arts. Moreover, the generated adversarial examples can be used to successfully boost the robustness and performance of the victim model via adversarial training.

Towards Unsupervised Speech Synthesis
Alexander H. Liu, Cheng-I Lai and James R. Glass 08:00-09:00 (702 Clearwater)
In this paper, we introduce the first unsupervised speech synthesis system that can be built with a simple recipe. The framework is based on a recently developed unsupervised speech recognition system and an existing neural-based speech synthesis paradigm. With unpaired audio, unpaired text, and lexicon, our method enables speech synthesis without the need for human-labeled corpus. Our preliminary result shows the unsupervised model achieved similar performance to its supervised counterpart in human opinion score.

"Again, Dozens of Refugees Drowned": A Computational Study of Political Framing Evoked by Presuppositions
Qi Yu 08:00-09:00 (702 Clearwater)
Earlier NLP studies on framing in political discourse have focused heavily on shallow classification of issue framing, while framing effect

arising from pragmatic cues remains neglected. We put forward this latter type of framing as "pragmatic framing". To bridge this gap, we take presupposition-triggering adverbs such as 'again' as a study case, and quantitatively investigate how different German newspapers use them to covertly evoke different attitudinal subtexts in their report on the event "European Refugee Crisis" (2014-2018). Our study demonstrates the crucial role of presuppositions in framing, and emphasizes the necessity of more attention on pragmatic framing in the research of automated framing detection.

Impact of Training Instance Selection on Domain-Specific Entity Extraction using BERT
Eileen Salhofer, Xing Lan Liu and Roman Kern 08:00-09:00 (702 Clearwater)
State of the art performances for entity extraction tasks are achieved by supervised learning, specifically, by fine-tuning pretrained language models such as BERT. As a result, annotating application specific data is the first step in many use cases. However, no practical guidelines are available for annotation requirements. This work supports practitioners by empirically answering the frequently asked questions (1) how many training samples to annotate? (2) which examples to annotate? We found that BERT achieves up to 80% F1 when fine-tuned on only 70 training examples, especially on biomedical domain. The key features for guiding the selection of high performing training instances are identified to be pseudo-perplexity and sentence-length. The best training dataset constructed using our proposed selection strategy shows F1 score that is equivalent to a random selection with twice the sample size. The requirement of only a small number of training data implies cheaper implementations and opens door to wider range of applications.

Investigating the effectiveness of various speaker embeddings for multi-speaker end-to-end speech synthesis system using small-sized speech data
Sheng-Yao Wang and Yi-Chin Huang 08:00-09:00 (702 Clearwater)
In this paper, we investigated the effectiveness of incorporating various speaker embeddings into an end-to-end speech synthesis system, for generating a unseen speaker's voice with small-sized speech data. To do so, we adopted learned speaker embeddings from various tasks, such as voice conversion and speaker verification. By combining the speaker embeddings using additive attention mechanism to an autoregressive-based speech synthesis framework, we could evaluate the performance of these embedding methods. To further enhance the speaker similarity and speech quality, the post-net for the output spectrogram sequence is replaced by a post-filter network. Experimental results showed that the proposed speech synthesis system with speaker embedding is capable of generating fluent arbitrary speech utterances of a unseen speaker with only few speech utterances. Besides, the post-filter network is helpful for enhancing the speaker similarity and speech naturalness of the output speech.

Text Style Transfer for Bias Mitigation using Masked Language Modeling
Ewoenam Kwaku Tokpo and Toon Calders 08:00-09:00 (702 Clearwater)
It is well known that textual data on the internet and other digital platforms contain significant levels of bias and stereotypes. Various research findings have concluded that biased texts have significant effects on target demographic groups. For instance, masculine-worded job advertisements tend to be less appealing to female applicants. In this paper, we present a text-style transfer model that can be trained on non-parallel data and be used to automatically mitigate bias in textual data. Our style transfer model improves on the limitations of many existing text style transfer techniques such as the loss of content information. Our model solves such issues by combining latent content encoding with explicit keyword replacement. We will show that this technique produces better content preservation whilst maintaining good style transfer accuracy.

Eliciting Complex Relational Knowledge From Masked Language Models
Arun Sundaresan, Ming Hsu and Zhihao Zhang 08:00-09:00 (702 Clearwater)
We present results from a series of experiments that probe the ability of masked language models (MLMs), such as BERT and RoBERTa, to respond to general knowledge questions that do not have a single correct answer. Our investigation leverages the semantic fluency task from cognitive science, in which a variable number of exemplars from a semantic category (e.g., fruits) need to be produced in a specific order. It allows us to evaluate what MLMs know about common categories and their members, a representative type of one-to-many relational knowledge, and how they organize and query such knowledge. We developed incremental cloze tasks that reflect serial knowledge search, and show that MLMs, especially RoBERTa, are able to generate semantic fluency responses that strongly resemble responses from human subjects in both their content and dynamics. These findings contribute to the literature on whether and how masked language models can be used as knowledge bases, and also provide novel insights on their knowledge structure.

Differentially Private Instance Encoding against Privacy Attacks
Shangyu Xie and Yuan Hong 08:00-09:00 (702 Clearwater)
TextHide was recently proposed to protect the training data via instance encoding in natural language domain. Due to the lack of theoretic privacy guarantee, such instance encoding scheme has been shown to be vulnerable against privacy attacks, e.g., reconstruction attack. To address such limitation, we revise the instance encoding scheme with differential privacy and thus provide a provable guarantee against privacy attacks. The experimental results also show that the proposed scheme can defend against privacy attacks while ensuring learning utility (as a trade-off).

Methods for Estimating and Improving Robustness of Language Models
Michal Stefanik 08:00-09:00 (702 Clearwater)
Despite their outstanding performance, large language models (LLMs) suffer notorious flaws related to their preference for shallow textual relations over full semantic complexity of the problem. This proposal investigates a common denominator of this problem in their weak ability to generalise outside of the training domain. We survey diverse research directions providing estimations of model generalisation ability and find that incorporating some of these measures in the training objectives leads to enhanced distributional robustness of neural models. Based on these findings, we present future research directions enhancing the robustness of LLMs.

Static and Dynamic Speaker Modeling based on Graph Neural Network for Emotion Recognition in Conversation
Prakhar Saxena, Yin Jou Huang and Sadao Kurohashi 08:00-09:00 (702 Clearwater)
Each person has a unique personality which affects how they feel and convey emotions. Hence, speaker modeling is important for the task of emotion recognition in conversation (ERC). In this paper, we propose a novel graph-based ERC model which considers both conversational context and speaker personality. We model the internal state of the speaker (personality) as Static and Dynamic speaker state, where the Dynamic speaker state is modeled with a graph neural network based encoder. Experiments on benchmark dataset shows the effectiveness of our model. Our model outperforms baseline and other graph-based methods. Analysis of results also show the importance of explicit speaker modeling.

Simulating Feature Structures with Simple Types
Valentin D. Richard 08:00-09:00 (702 Clearwater)
Feature structures have been several times considered to enrich categorial grammars in order to build fine-grained grammars. Most attempts to unify both frameworks either model categorial types as feature structures or add feature structures on top of categorial types. We pursue a different approach: using feature structure as categorial atomic types. In this article, we present a procedure to create, from a simplified HPSG

grammar, an equivalent abstract categorial grammar (ACG). We represent a feature structure by the enumeration of its totally well-typed upper bounds, so that unification can be simulated as intersection. We implement this idea as a meta-ACG preprocessor.

Virtual Poster Q&A Session 1

08:00-09:00 (702 Clearwater)

Explicit Use of Topicality in Dialogue Response Generation
Takumi Yoshikoshi, Hayato Atarashi, Takashi Kodama and Sadao Kurohashi 08:00-09:00 (702 Clearwater)
The current chat dialogue systems implicitly consider the topic given the context, but not explicitly. As a result, these systems often generate inconsistent responses with the topic of the moment. In this study, we propose a dialogue system that responds appropriately following the topic by selecting the entity with the highest "topicality." In topicality estimation, the model is trained through self-supervised learning that regards entities that appear in both context and response as the topic entities. In response generation, the model is trained to generate topic-relevant responses based on the estimated topicality. Experimental results show that our proposed system can follow the topic more than the existing dialogue system that considers only the context.

Automating Human Evaluation of Dialogue Systems
Sujan Reddy A 08:00-09:00 (702 Clearwater)
Automated metrics to evaluate dialogue systems like BLEU, METEOR, etc., weakly correlate with human judgments. Thus, human evaluation is often used to supplement these metrics for system evaluation. However, human evaluation is time-consuming as well as expensive. This paper provides an alternative approach to human evaluation with respect to three aspects: naturalness, informativeness, and quality in dialogue systems. I propose an approach based on fine-tuning the BERT model with three prediction heads, to predict whether the system-generated output is natural, fluent, and informative. I observe that the proposed model achieves an average accuracy of around 77% over these 3 labels. I also design a baseline approach that uses three different BERT models to make the predictions. Based on experimental analysis, I find that using a shared model to compute the three labels performs better than three separate models.

Generating Repetitions with Appropriate Repeated Words
Toshiki Kawamoto, Hidetaka Kamigaito, Kotaro Funakoshi and Manabu Okumura 08:00-09:00 (702 Clearwater)
A repetition is a response that repeats words in the previous speaker's utterance in a dialogue. Repetitions are essential in communication to build trust with others, as investigated in linguistic studies. In this work, we focus on repetition generation. To the best of our knowledge, this is the first neural approach to address repetition generation. We propose Weighted Label Smoothing, a smoothing method for explicitly learning which words to repeat during fine-tuning, and a repetition scoring method that can output more appropriate repetitions during decoding. We conducted automatic and human evaluations involving applying these methods to the pre-trained language model T5 for generating repetitions. The experimental results indicate that our methods outperformed baselines in both evaluations.

EmpHi: Generating Empathetic Responses with Human-like Intents
Mao Yan Chen, Siheng Li and Yujiu Yang 08:00-09:00 (702 Clearwater)
In empathetic conversations, humans express their empathy to others with empathetic intents. However, most existing empathetic conversational methods suffer from a lack of empathetic intents, which leads to monotonous empathy. To address the bias of the empathetic intents distribution between empathetic dialogue models and humans, we propose a novel model to generate empathetic responses with human-consistent empathetic intents, EmpHi for short. Precisely, EmpHi learns the distribution of potential empathetic intents with a discrete latent variable, then combines both implicit and explicit intent representation to generate responses with various empathetic intents. Experiments show that EmpHi outperforms state-of-the-art models in terms of empathy, relevance, and diversity on both automatic and human evaluation. Moreover, the case studies demonstrate the high interpretability and outstanding performance of our model.

Towards a Progression-Aware Autonomous Dialogue Agent
Abraham Sanders, Tomek Strzalkowski, Mei Si, Albert Chang, Deepanshu Dey, Jonas Braasch and Dakuo Wang 08:00-09:00 (702 Clearwater)
Recent advances in large-scale language modeling and generation have enabled the creation of dialogue agents that exhibit human-like responses in a wide range of conversational scenarios spanning a diverse set of tasks, from general chit-chat to focused goal-oriented discourse. While these agents excel at generating high-quality responses that are relevant to prior context, they suffer from a lack of awareness of the overall direction in which the conversation is headed, and the likelihood of task success inherent therein. Thus, we propose a framework in which dialogue agents can evaluate the progression of a conversation toward or away from desired outcomes, and use this signal to inform planning for subsequent responses. Our framework is composed of three key elements: (1) the notion of a "global" dialogue state (GDS) space, (2) a task-specific progression function (PF) computed in terms of a conversation's trajectory through this space, and (3) a planning mechanism based on dialogue rollouts by which an agent may use progression signals to select its next response.

Representation Learning for Conversational Data using Discourse Mutual Information Maximization
Bishal Santra, Sumegh Roychowdhury, Aishik Mandal, Vasu Gurram, Atharva Naik, Manish Gupta and Pawan Goyal 08:00-09:00 (702 Clearwater)
Although many pretrained models exist for text or images, there have been relatively fewer attempts to train representations specifically for dialog understanding. Prior works usually relied on finetuned representations based on generic text representation models like BERT or GPT-2. But such language modeling pretraining objectives do not take the structural information of conversational text into consideration. Although generative dialog models can learn structural features too, we argue that the structure-unaware word-by-word generation is not suitable for effective conversation modeling. We empirically demonstrate that such representations do not perform consistently across various dialog understanding tasks. Hence, we propose a structure-aware Mutual Information based loss-function DMI (Discourse Mutual Information) for training dialog-representation models, that additionally captures the inherent uncertainty in response prediction. Extensive evaluation on nine diverse dialog modeling tasks shows that our proposed DMI-based models outperform strong baselines by significant margins.

D2U: Distance-to-Uniform Learning for Out-of-Scope Detection
Eyup Halit Yilmaz and Cagri Toraman 08:00-09:00 (702 Clearwater)
Supervised training with cross-entropy loss implicitly forces models to produce probability distributions that follow a discrete delta distribution. Model predictions in test time are expected to be similar to delta distributions if the classifier determines the class of an input correctly. However, the shape of the predicted probability distribution can become similar to the uniform distribution when the model cannot infer properly. We exploit this observation for detecting out-of-scope (OOS) utterances in conversational systems. Specifically, we propose a zero-shot post-processing step, called Distance-to-Uniform (D2U), exploiting not only the classification confidence score, but the shape of the entire

output distribution. We later combine it with a learning procedure that uses D2U for loss calculation in the supervised setup. We conduct experiments using six publicly available datasets. Experimental results show that the performance of OOS detection is improved with our post-processing when there is no OOS training data, as well as with D2U learning procedure when OOS training data is available.

Building a Role Specified Open-Domain Dialogue System Leveraging Large-Scale Language Models
Sanghwan Bae, Donghyun Kwak, Sungdong Kim, Donghoon Ham, Soyoung Kang, Sang-Woo Lee and Woomyoung Park 08:00-09:00 (702 Clearwater)
Recent open-domain dialogue models have brought numerous breakthroughs. However, building a chat system is not scalable since it often requires a considerable volume of human-human dialogue data, especially when enforcing features such as persona, style, or safety. In this work, we study the challenge of imposing roles on open-domain dialogue systems, with the goal of making the systems maintain consistent roles while conversing naturally with humans. To accomplish this, the system must satisfy a role specification that includes certain conditions on the stated features as well as a system policy on whether or not certain types of utterances are allowed. For this, we propose an efficient data collection framework leveraging in-context few-shot learning of large-scale language models for building role-satisfying dialogue dataset from scratch. We then compare various architectures for open-domain dialogue systems in terms of meeting role specifications while maintaining conversational abilities. Automatic and human evaluations show that our models return few out-of-bounds utterances, keeping competitive performance on general metrics. We release a Korean dialogue dataset we built for further research.

Revisit Overconfidence for OOD Detection: Reassigned Contrastive Learning with Adaptive Class-dependent Threshold
Yanan Wu, Keqing He, Yuanmeng Yan, QiXiang Gao, Zhiyuan Zeng, Fujia Zheng, Lulu Zhao, Huixing Jiang, Wei Wu and Weiran Xu 08:00-09:00 (702 Clearwater)
Detecting Out-of-Domain (OOD) or unknown intents from user queries is essential in a task-oriented dialog system. A key challenge of OOD detection is the overconfidence of neural models. In this paper, we comprehensively analyze overconfidence and classify it into two perspectives: over-confident OOD and in-domain (IND). Then according to intrinsic reasons, we respectively propose a novel reassigned contrastive learning (RCL) to discriminate IND intents for over-confident OOD and an adaptive class-dependent local threshold mechanism to separate similar IND and OOD intents for over-confident IND. Experiments and analyses show the effectiveness of our proposed method for both aspects of overconfidence issues.

AISFG: Abundant Information Slot Filling Generator
Yang Yan, Junda Ye, Zhongbao Zhang and Liwen Wang 08:00-09:00 (702 Clearwater)
As an essential component of task-oriented dialogue systems, slot filling requires enormous labeled training data in a certain domain. However, in most cases, there is little or no target domain training data is available in the training stage. Thus, cross-domain slot filling has to cope with the data scarcity problem by zero/few-shot learning. Previous researches on zero/few-shot cross-domain slot filling focus on slot descriptions and examples while ignoring the slot type ambiguity and example ambiguity issues. To address these problems, we propose Abundant Information Slot Filling Generator (AISFG), a generative model with a novel query template that incorporates domain descriptions, slot descriptions, and examples with context. Experimental results show that our model outperforms state-of-the-art approaches in zero/few-shot slot filling task.

Mining Clues from Incomplete Utterance: A Query-enhanced Network for Incomplete Utterance Rewriting
Shuzheng Si, Shuang Zeng and Baobao Chang 08:00-09:00 (702 Clearwater)
Incomplete utterance rewriting has recently raised wide attention. However, previous works do not consider the semantic structural information between incomplete utterance and rewritten utterance or model the semantic structure implicitly and insufficiently. To address this problem, we propose a QUEry-Enhanced Network(QUEEN) to solve this problem. Firstly, our proposed query template explicitly brings guided semantic structural knowledge between the incomplete utterance and the rewritten utterance making model perceive where to refer back to or recover omitted tokens. Then, we adopt a fast and effective edit operation scoring network to model the relation between two tokens. Benefiting from extra information and the well-designed network, QUEEN achieves state-of-the-art performance on several public datasets.

Meet Your Favorite Character: Open-domain Chatbot Mimicking Fictional Characters with only a Few Utterances
Seungju Han, Beomsu Kim, Jin Yong Yoo, Seokjun Seo, Sangbum Kim, Enkhbayar Erdenee and Buru Chang 08:00-09:00 (702 Clearwater)
In this paper, we consider mimicking fictional characters as a promising direction for building engaging conversation models. To this end, we present a new practical task where only a few utterances of each fictional character are available to generate responses mimicking them. Furthermore, we propose a new method named Pseudo Dialog Prompting (PDP) that generates responses by leveraging the power of large-scale language models with prompts containing the target character's utterances. To better reflect the style of the character, PDP builds the prompts in the form of dialog that includes the character's utterances as dialog history. Since only utterances of the characters are available in the proposed task, PDP matches each utterance with an appropriate pseudo-context from a predefined set of context candidates using a retrieval model. Through human and automatic evaluation, we show that PDP generates responses that better reflect the style of fictional characters than baseline methods.

Learning to Execute Actions or Ask Clarification Questions
Zhengxiang Shi, Yue Feng and Aldo Lipani 08:00-09:00 (702 Clearwater)
Collaborative tasks are ubiquitous activities where a form of communication is required in order to reach a joint goal. Collaborative building is one of such tasks. We wish to develop an intelligent builder agent in a simulated building environment (Minecraft) that can build whatever users wish to build by just talking to the agent. In order to achieve this goal, such agents need to be able to take the initiative by asking clarification questions when further information is needed. Existing works on Minecraft Corpus Dataset only learn to execute instructions neglecting the importance of asking for clarifications. In this paper, we extend the Minecraft Corpus Dataset by annotating all builder utterances into eight types, including clarification questions, and propose a new builder agent model capable of determining when to ask or execute instructions. Experimental results show that our model achieves state-of-the-art performance on the collaborative building task with a substantial improvement. We also define two new tasks, the learning to ask task and the joint learning task. The latter consists of solving both collaborating building and learning to ask tasks jointly.

BORT: Back and Denoising Reconstruction for End-to-End Task-Oriented Dialog
Haipeng Sun, Junwei Bao, Youzheng Wu and Xiaodong He 08:00-09:00 (702 Clearwater)
A typical end-to-end task-oriented dialog system transfers context into dialog state, and upon which generates a response, which usually faces the problem of error propagation from both previously generated inaccurate dialog states and responses, especially in low-resource scenarios. To alleviate these issues, we propose BORT, a back and denoising reconstruction approach for end-to-end task-oriented dialog system. Squarely, to improve the accuracy of dialog states, back reconstruction is used to reconstruct the original input context from the generated dialog states since inaccurate dialog states cannot recover the corresponding input context. To enhance the denoising capability of the model to reduce the impact of error propagation, denoising reconstruction is used to reconstruct the corrupted dialog state and response. Extensive experiments conducted on MultiWOZ 2.0 and CamRest676 show the effectiveness of BORT. Furthermore, BORT demonstrates its advanced

capabilities in the zero-shot domain and low-resource scenarios.

Am I Me or You? State-of-the-Art Dialogue Models Cannot Maintain an Identity
Kurt Shuster, Jack Urbanek, Arthur Szlam and Jason E Weston 08:00-09:00 (702 Clearwater)
State-of-the-art dialogue models still often stumble with regards to factual accuracy and self-contradiction. Anecdotally, they have been observed to fail to maintain character identity throughout discourse; and more specifically, may take on the role of their interlocutor. In this work we formalize and quantify this deficiency, and show experimentally through human evaluations that this is indeed a problem. In contrast, we show that discriminative models trained specifically to recognize who is speaking can perform well; and further, these can be used as automated metrics. Finally, we evaluate a wide variety of mitigation methods, including changes to model architecture, training protocol, and decoding strategy. Our best models reduce mistaken identity issues by nearly 65% according to human annotators, while simultaneously improving engagingness. Despite these results, we find that maintaining character identity still remains a challenging problem.

DialoKG: Knowledge-Structure Aware Task-Oriented Dialogue Generation
Md Rashad Al Hasan Rony, Ricardo Usbeck and Jens Lehmann 08:00-09:00 (702 Clearwater)
Task-oriented dialogue generation is challenging since the underlying knowledge is often dynamic and effectively incorporating knowledge into the learning process is hard. It is particularly challenging to generate both human-like and informative responses in this setting. Recent research primarily focused on various knowledge distillation methods where the underlying relationship between the facts in a knowledge base is not effectively captured. In this paper, we go one step further and demonstrate how the structural information of a knowledge graph can improve the system's inference capabilities. Specifically, we propose DialoKG, a novel task-oriented dialogue system that effectively incorporates knowledge into a language model. Our proposed system views relational knowledge as a knowledge graph and introduces (1) a structure-aware knowledge embedding technique, and (2) a knowledge graph-weighted attention masking strategy to facilitate the system selecting relevant information during the dialogue generation. An empirical evaluation demonstrates the effectiveness of DialoKG over state-of-the-art methods on several standard benchmark datasets.

Relation-Specific Attentions over Entity Mentions for Enhanced Document-Level Relation Extraction
Jiaxin Yu, Deqing Yang and Shuyu Tian 08:00-09:00 (702 Clearwater)
Compared with traditional sentence-level relation extraction, document-level relation extraction is a more challenging task where an entity in a document may be mentioned multiple times and associated with multiple relations. However, most methods of document-level relation extraction do not distinguish between mention-level features and entity-level features, and just apply simple pooling operation for aggregating mention-level features into entity-level features. As a result, the distinct semantics between the different mentions of an entity are overlooked. To address this problem, we propose RSMAN in this paper which performs selective attentions over different entity mentions with respect to candidate relations. In this manner, the flexible and relation-specific representations of entities are obtained which indeed benefit relation classification. Our extensive experiments upon two benchmark datasets show that our RSMAN can bring significant improvements for some backbone models to achieve state-of-the-art performance, especially when an entity have multiple mentions in the document.

Hero-Gang Neural Model For Named Entity Recognition
Jinpeng Hu, Yaling Shen, Yang Liu, Xiang Wan and Tsung-Hui Chang 08:00-09:00 (702 Clearwater)
Named entity recognition (NER) is a fundamental and important task in NLP, aiming at identifying named entities (NEs) from free text. Recently, since the multi-head attention mechanism applied in the Transformer model can effectively capture longer contextual information, Transformer-based models have become the mainstream methods and have achieved significant performance in this task. Unfortunately, although these models can capture effective global context information, they are still limited in the local feature and position information extraction, which is critical in NER. In this paper, to address this limitation, we propose a novel Hero-Gang Neural structure (HGN), including the Hero and Gang module, to leverage both global and local information to promote NER. Specifically, the Hero module is composed of a Transformer-based encoder to maintain the advantage of the self-attention mechanism, and the Gang module utilizes a multi-window recurrent module to extract local features and position information under the guidance of the Hero module. Afterward, the proposed multi-window attention effectively combines global information and multiple local features for predicting entity labels. Experimental results on several benchmark datasets demonstrate the effectiveness of our proposed model.

Modal Dependency Parsing via Language Model Priming
Jiarui Yao, Nianwen Xue and Bonan Min 08:00-09:00 (702 Clearwater)
The task of modal dependency parsing aims to parse a text into its modal dependency structure, which is a representation for the factuality of events in the text. We design a modal dependency parser that is based on priming pre-trained language models, and evaluate the parser on two data sets. Compared to baselines, we show an improvement of 2.6% in F-score for English and 4.6% for Chinese. To the best of our knowledge, this is also the first work on Chinese modal dependency parsing.

Document-Level Event Argument Extraction by Leveraging Redundant Information and Closed Boundary Loss
Hanzhang Zhou and Kezhi Mao 08:00-09:00 (702 Clearwater)
In document-level event argument extraction, an argument is likely to appear multiple times in different expressions in the document. The redundancy of arguments underlying multiple sentences is beneficial but is often overlooked. In addition, in event argument extraction, most entities are regarded as class “others", i.e. Universum class, which is defined as a collection of samples that do not belong to any class of interest. Universum class is composed of heterogeneous entities without typical common features. Classifiers trained by cross entropy loss could easily misclassify the Universum class because of their open decision boundary. In this paper, to make use of redundant event information underlying a document, we build an entity coreference graph with the graph2token module to produce a comprehensive and coreference-aware representation for every entity and then build an entity summary graph to merge the multiple extraction results. To better classify Universum class, we propose a new loss function to build classifiers with closed boundaries. Experimental results show that our model outperforms the previous state-of-the-art models by 3.35% in F1-score.

Global Entity Disambiguation with BERT
Ikuya Yamada, Koki Washio, Hiroyuki Shindo and Yuji Matsumoto 08:00-09:00 (702 Clearwater)
We propose a global entity disambiguation (ED) model based on BERT. To capture global contextual information for ED, our model treats not only words but also entities as input tokens, and solves the task by sequentially resolving mentions to their referent entities and using resolved entities as inputs at each step. We train the model using a large entity-annotated corpus obtained from Wikipedia. We achieve new state-of-the-art results on five standard ED datasets: AIDA-CoNLL, MSNBC, AQUAINT, ACE2004, and WNED-WIKI. The source code and model checkpoint are available at https://github.com/studio-ousia/luke.

Does it Really Generalize Well on Unseen Data? Systematic Evaluation of Relational Triple Extraction Methods
Juhyuk Lee, Min-Joong Lee, June Yong Yang and Eunho Yang 08:00-09:00 (702 Clearwater)
The ability to extract entities and their relations from unstructured text is essential for the automated maintenance of large-scale knowledge graphs. To keep a knowledge graph up-to-date, an extractor needs not only the ability to recall the triples it encountered during training, but

also the ability to extract the new triples from the context that it has never seen before. In this paper, we show that although existing extraction models are able to easily memorize and recall already seen triples, they cannot generalize effectively for unseen triples. This alarming observation was previously unknown due to the composition of the test sets of the go-to benchmark datasets, which turns out to contain only 2% unseen data, rendering them incapable to measure the generalization performance. To separately measure the generalization performance from the memorization performance, we emphasize unseen data by rearranging datasets, sifting out training instances, or augmenting test sets. In addition to that, we present a simple yet effective augmentation technique to promote generalization of existing extraction models, and experimentally confirm that the proposed method can significantly increase the generalization performance of existing models.

Generative Biomedical Entity Linking via Knowledge Base-Guided Pre-training and Synonyms-Aware Fine-tuning
Hongyi Yuan, Zheng Yuan and Sheng Yu 08:00-09:00 (702 Clearwater)
Entities lie in the heart of biomedical natural language understanding, and the biomedical entity linking (EL) task remains challenging due to the fine-grained and diversiform concept names. Generative methods achieve remarkable performances in general domain EL with less memory usage while requiring expensive pre-training. Previous biomedical EL methods leverage synonyms from knowledge bases (KB) which is not trivial to inject into a generative method. In this work, we use a generative approach to model biomedical EL and propose to inject synonyms knowledge in it. We propose KB-guided pre-training by constructing synthetic samples with synonyms and definitions from KB and require the model to recover concept names. We also propose synonyms-aware fine-tuning to select concept names for training, and propose decoder prompt and multi-synonyms constrained prefix tree for inference. Our method achieves state-of-the-art results on several biomedical EL tasks without candidate selection which displays the effectiveness of proposed pre-training and fine-tuning strategies. The source code is available at `https://github.com/Yuanhy1997/GenBioEL`.

RAAT: Relation-Augmented Attention Transformer for Relation Modeling in Document-Level Event Extraction
Yuan Liang, Zhuoxuan Jiang, di Yin and Bo Ren 08:00-09:00 (702 Clearwater)
In document-level event extraction (DEE) task, event arguments always scatter across sentences (across-sentence issue) and multiple events may lie in one document (multi-event issue). In this paper, we argue that the relation information of event arguments is of great significance for addressing the above two issues, and propose a new DEE framework which can model the relation dependencies, called Relation-augmented Document-level Event Extraction (ReDEE). More specifically, this framework features a novel and tailored transformer, named as Relation-augmented Attention Transformer (RAAT). RAAT is scalable to capture multi-scale and multi-amount argument relations. To further leverage relation information, we introduce a separate event relation prediction task and adopt multi-task learning method to explicitly enhance event extraction performance. Extensive experiments demonstrate the effectiveness of the proposed method, which can achieve state-of-the-art performance on two public datasets.Our code is available at https://github.com/TencentYoutuResearch/RAAT.

Improving Few-Shot Relation Classification by Prototypical Representation Learning with Definition Text
Li Zhenzhen, Yuyang Zhang, Jian-Yun Nie and Dongsheng Li 08:00-09:00 (702 Clearwater)
Few-shot relation classification is difficult because the few instances available may not represent well the relation patterns. Some existing approaches explored extra information such as relation definition, in addition to the instances, to learn a better relation representation. However, the encoding of the extra information has been performed independently from the labeled instances. In this paper, we propose to learn a prototype encoder from relation definition in a way that is useful for relation instance classification. To this end, we use a joint training approach to train both a prototype encoder from definition and an instance encoder. Extensive experiments on several datasets demonstrate the effectiveness and usefulness of our prototype encoder from definition text, enabling us to outperform state-of-the-art approaches.

Dependency Position Encoding for Relation Extraction
Qiushi Guo, Xin Wang and Dehong Gao 08:00-09:00 (702 Clearwater)
Leveraging the dependency tree of the input sentence is able to improve the model performance for relation extraction. A challenging issue is how to remove confusions from the tree. Efforts have been made to utilize the dependency connections between words to selectively emphasize target-relevant information. However, these approaches are limited in focusing on exploiting dependency types. In this paper, we propose dependency position encoding (DPE), an efficient way of incorporating both dependency connections and dependency types into the self-attention mechanism to distinguish the importance of different word dependencies for the task. In contrast to previous studies that process input sentence and dependency information in separate streams, DPE can be seamlessly incorporated into the Transformer and makes it possible to use an one-stream scheme to extract relations between entity pairs. Extensive experiments show that models with our DPE significantly outperform the previous methods on SemEval 2010 Task 8, KBP37, and TACRED.

XLTime: A Cross-Lingual Knowledge Transfer Framework for Temporal Expression Extraction
*Yuwei Cao, William Groves, Tanay Kumar Saha, Joel R. Tetreault, Alejandro Jaimes, Hao Peng and Philip S. Yu*08:00-09:00 (702 Clearwater)
Temporal Expression Extraction (TEE) is essential for understanding time in natural language. It has applications in Natural Language Processing (NLP) tasks such as question answering, information retrieval, and causal inference. To date, work in this area has mostly focused on English as there is a scarcity of labeled data for other languages. We propose XLTime, a novel framework for multilingual TEE. XLTime works on top of pre-trained language models and leverages multi-task learning to prompt cross-language knowledge transfer both from English and within the non-English languages. XLTime alleviates problems caused by a shortage of data in the target language. We apply XLTime with different language models and show that it outperforms the previous automatic SOTA methods on French, Spanish, Portuguese, and Basque, by large margins. XLTime also closes the gap considerably on the handcrafted HeidelTime method.

A Label-Aware Autoregressive Framework for Cross-Domain NER
Jinpeng Hu, He Zhao, Dan Dan Guo, Xiang Wan and Tsung-Hui Chang 08:00-09:00 (702 Clearwater)
Cross-domain named entity recognition (NER) aims to borrow the entity information from the source domain to help the entity recognition in the target domain with limited labeled data. Despite the promising performance of existing approaches, most of them focus on reducing the discrepancy of token representation between source and target domains, while the transfer of the valuable label information is often not explicitly considered or even ignored. Therefore, we propose a novel autoregressive framework to advance cross-domain NER by first enhancing the relationship between labels and tokens and then further improving the transferability of label information. Specifically, we associate each label with an embedding vector, and for each token, we utilize a bidirectional LSTM (Bi-LSTM) to encode the labels of its previous tokens for modeling internal context information and label dependence. Afterward, we propose a Bi-Attention module that merges the token representation from a pre-trained model and the label features from the Bi-LSTM as the label-aware information, which is concatenated to the token representation to facilitate cross-domain NER. In doing so, label information contained in the embedding vectors can be effectively transferred to the target domain, and Bi-LSTM can further model the label relationship among different domains by pre-train and then fine-tune setting. Experimental results on several datasets confirm the effectiveness of our model, where our model achieves significant improvements over the state of the arts.

Learning Discriminative Representations for Open Relation Extraction with Instance Ranking and Label Calibration
Shusen Wang, Bin Duan, Yanan Wu and Yajing Xu 08:00-09:00 (702 Clearwater)
Open relation extraction is the task to extract relational facts without pre-defined relation types from open-domain corpora. However, since

there are some hard or semi-hard instances sharing similar context and entity information but belonging to different underlying relation, current OpenRE methods always cluster them into the same relation type. In this paper, we propose a novel method based on Instance Ranking and Label Calibration strategies (IRLC) to learn discriminative representations for open relation extraction. Due to lacking the original instance label, we provide three surrogate strategies to generate the positive, hard negative, and semi-hard negative instances for the original instance. Instance ranking aims to refine the relational feature space by pushing the hard and semi-hard negative instances apart from the original instance with different margins and pulling the original instance and its positive instance together. To refine the cluster probability distributions of these instances, we introduce a label calibration strategy to model the constraint relationship between instances. Experimental results on two public datasets demonstrate that our proposed method can significantly outperform the previous state-of-the-art methods.

RCL: Relation Contrastive Learning for Zero-Shot Relation Extraction
Shusen Wang, Bosen Zhang, Yajing Xu, Yanan Wu and Bo Xiao 08:00-09:00 (702 Clearwater)
Zero-shot relation extraction aims to identify novel relations which cannot be observed at the training stage. However, it still faces some challenges since the unseen relations of instances are similar or the input sentences have similar entities, the unseen relation representations from different categories tend to overlap and lead to errors. In this paper, we propose a novel Relation Contrastive Learning framework (RCL) to mitigate above two types of similar problems: Similar Relations and Similar Entities. By jointly optimizing a contrastive instance loss with a relation classification loss on seen relations, RCL can learn subtle difference between instances and achieve better separation between different relation categories in the representation space simultaneously. Especially in contrastive instance learning, the dropout noise as data augmentation is adopted to amplify the semantic difference between similar instances without breaking relation representation, so as to promote model to learn more effective representations. Experiments conducted on two well-known datasets show that RCL can significantly outperform previous state-of-the-art methods. Moreover, if the seen relations are insufficient, RCL can also obtain comparable results with the model trained on the full training set, showing the robustness of our approach.

Zero-Shot Event Detection Based on Ordered Contrastive Learning and Prompt-Based Prediction
Senhui Zhang, Tao Ji, Wendi Ji and Xiaoling Wang 08:00-09:00 (702 Clearwater)
Event detection is a classic natural language processing task. However, the constantly emerging new events make supervised methods not applicable to unseen types. Previous zero-shot event detection methods either require predefined event types as heuristic rules or resort to external semantic analyzing tools. To overcome this weakness, we propose an end-to-end framework named Zero-Shot Event Detection Based on Ordered Contrastive Learning and Prompt-Based Prediction (ZEOP). By creatively introducing multiple contrastive samples with ordered similarities, the encoder can learn event representations from both instance-level and class-level, which makes the distinctions between different unseen types more significant. Meanwhile, we utilize the prompt-based prediction to identify trigger words without relying on external resources. Experiments demonstrate that our model detects events more effectively and accurately than state-of-the-art methods.

Minimally-Supervised Relation Induction from Pre-trained Language Model
Lu Sun, Yongliang Shen and Weiming Lu 08:00-09:00 (702 Clearwater)
Relation Induction is a very practical task in Natural Language Processing (NLP) area. In practical application scenarios, people want to induce more entity pairs having the same relation from only a few seed entity pairs. Thus, instead of the laborious supervised setting, in this paper, we focus on the minimally-supervised setting where only a couple of seed entity pairs per relation are provided. Although the conventional relation induction methods have made some success, their performance depends heavily on the quality of word embeddings. The great success of Pre-trained Language Models, such as BERT, changes the NLP area a lot, and they are proven to be able to better capture relation knowledge. In this paper, we propose a novel method to induce relation with BERT under the minimally-supervised setting. Specifically, we firstly extract proper templates from the corpus by using the mask-prediction task in BERT to build pseudo-sentences as the context of entity pairs. Then we use BERT attention weights to better represent the pseudo-sentences. In addition, We also use the IntegratedGradient of entity pairs to iteratively select better templates further. Finally, with the high-quality pseudo-sentences, we can train a better classifier for relation induction. Experiments onGoogle Analogy Test Sets (GATS), Bigger Analogy TestSet (BATS) and DiffVec demonstrate that our proposed method achieves state-of-the-art performance.

Learn from Relation Information: Towards Prototype Representation Rectification for Few-Shot Relation Extraction
Yang Liu, Jinpeng Hu, Xiang Wan and Tsung-Hui Chang 08:00-09:00 (702 Clearwater)
Few-shot Relation Extraction refers to fast adaptation to novel relation classes with few samples through training on the known relation classes. Most existing methods focus on implicitly introducing relation information (i.e., relation label or relation description) to constrain the prototype representation learning, such as contrastive learning, graphs, and specifically designed attentions, which may bring useless and even harmful parameters. Besides, these approaches are limited in handing outlier samples far away from the class center due to the weakly implicit constraint. In this paper, we propose an effective and parameter-less Prototype Rectification Method (PRM) to promote few-shot relation extraction, where we utilize a prototype rectification module to rectify original prototypes explicitly by the relation information. Specifically, PRM is composed of two gate mechanisms. One gate decides how much of the original prototype remains, and another one updates the remained prototype with relation information. In doing so, better and stabler global relation information can be captured for guiding prototype representations, and thus PRM can robustly deal with outliers. Moreover, we also extend PRM to both none-of-the-above (NOTA) and domain adaptation scenarios. Experimental results on FewRel 1.0 and 2.0 datasets demonstrate the effectiveness of our proposed method, which achieves state-of-the-art performance.

Delving Deep into Regularity: A Simple but Effective Method for Chinese Named Entity Recognition
Yingjie Gu, Xiaoye Qu, Zhefeng Wang, Yi Zheng, Baoxing Huai and Nicholas Jing Yuan 08:00-09:00 (702 Clearwater)
Recent years have witnessed the improving performance of Chinese Named Entity Recognition (NER) from proposing new frameworks or incorporating word lexicons. However, the inner composition of entity mentions in character-level Chinese NER has been rarely studied. Actually, most mentions of regular types have strong name regularity. For example, entities end with indicator words such as "公司 (company) " or “銀行 (bank)” usually belong to organization. In this paper, we propose a simple but effective method for investigating the regularity of entity spans in Chinese NER, dubbed as Regularity-Inspired reCOgnition Network (RICON). Specifically, the proposed model consists of two branches: a regularity-aware module and a regularity-agnostic module. The regularity-aware module captures the internal regularity of each span for better entity type prediction, while the regularity-agnostic module is employed to locate the boundary of entities and relieve the excessive attention to span regularity. An orthogonality space is further constructed to encourage two modules to extract different aspects of regularity features. To verify the effectiveness of our method, we conduct extensive experiments on three benchmark datasets and a practical medical dataset. The experimental results show that our RICON significantly outperforms previous state-of-the-art methods, including various lexicon-based methods.

Probe-Less Probing of BERT’s Layer-Wise Linguistic Knowledge with Masked Word Prediction
Tatsuya Aoyama and Nathan Schneider 08:00-09:00 (702 Clearwater)
The current study quantitatively (and qualitatively for an illustrative purpose) analyzes BERT’s layer-wise masked word prediction on an English corpus, and finds that (1) the layerwise localization of linguistic knowledge primarily shown in probing studies is replicated in a

behavior-based design and (2) that syntactic and semantic information is encoded at different layers for words of different syntactic categories. Hypothesizing that the above results are correlated with the number of likely potential candidates of the masked word prediction, we also investigate how the results differ for tokens within multiword expressions.

Same Neurons, Different Languages: Probing Morphosyntax in Multilingual Pre-trained Models
Karolina Stanczak, Edoardo Ponti, Lucas Torroba Hennigen, Ryan Cotterell and Isabelle Augenstein 08:00-09:00 (702 Clearwater)
The success of multilingual pre-trained models is underpinned by their ability to learn representations shared by multiple languages even in absence of any explicit supervision. However, it remains unclear how these models learn to generalise across languages. In this work, we conjecture that multilingual pre-trained models can derive language-universal abstractions about grammar. In particular, we investigate whether morphosyntactic information is encoded in the same subset of neurons in different languages. We conduct the first large-scale empirical study over 43 languages and 14 morphosyntactic categories with a state-of-the-art neuron-level probe. Our findings show that the cross-lingual overlap between neurons is significant, but its extent may vary across categories and depends on language proximity and pre-training data size.

How Conservative are Language Models? Adapting to the Introduction of Gender-Neutral Pronouns
Stephanie Brandl, Ruixiang Cui and Anders Søgaard 08:00-09:00 (702 Clearwater)
Gender-neutral pronouns have recently been introduced in many languages to a) include non-binary people and b) as a generic singular. Recent results from psycholinguistics suggest that gender-neutral pronouns (in Swedish) are not associated with human processing difficulties. This, we show, is in sharp contrast with automated processing. We show that gender-neutral pronouns in Danish, English, and Swedish are associated with higher perplexity, more dispersed attention patterns, and worse downstream performance. We argue that such conservativity in language models may limit widespread adoption of gender-neutral pronouns and must therefore be resolved.

Residue-Based Natural Language Adversarial Attack Detection
Vyas Raina and Mark Gales 08:00-09:00 (702 Clearwater)
Deep learning based systems are susceptible to adversarial attacks, where a small, imperceptible change at the input alters the model prediction. However, to date the majority of the approaches to detect these attacks have been designed for image processing systems. Many popular image adversarial detection approaches are able to identify adversarial examples from embedding feature spaces, whilst in the NLP domain existing state of the art detection approaches solely focus on input text features, without consideration of model embedding spaces. This work examines what differences result when porting these image designed strategies to Natural Language Processing (NLP) tasks - these detectors are found to not port over well. This is expected as NLP systems have a very different form of input: discrete and sequential in nature, rather than the continuous and fixed size inputs for images. As an equivalent model-focused NLP detection approach, this work proposes a simple sentence-embedding "residue" based detector to identify adversarial examples. On many tasks, it out-performs ported image domain detectors and recent state of the art NLP specific detectors.

Models In a Spelling Bee: Language Models Implicitly Learn the Character Composition of Tokens
Itay Itzhak and Omer Levy 08:00-09:00 (702 Clearwater)
Standard pretrained language models operate on sequences of subword tokens without direct access to the characters that compose each token's string representation. We probe the embedding layer of pretrained language models and show that models learn the internal character composition of whole word and subword tokens to a surprising extent, without ever seeing the characters coupled with the tokens. Our results show that the embedding layers of RoBERTa and GPT2 each hold enough information to accurately spell up to a third of the vocabulary and reach high character ngram overlap across all token types. We further test whether enriching subword models with character information can improve language modeling, and observe that this method has a near-identical learning curve as training without spelling-based enrichment. Overall, our results suggest that language modeling objectives incentivize the model to implicitly learn some notion of spelling, and that explicitly teaching the model how to spell does not appear to enhance its performance on such tasks.

Phrase-level Textual Adversarial Attack with Label Preservation
Yibin Lei, Yu Cao, Dianqi Li, Tianyi Zhou, Meng Fang and Mykola Pechenizkiy 08:00-09:00 (702 Clearwater)
Generating high-quality textual adversarial examples is critical for investigating the pitfalls of natural language processing (NLP) models and further promoting their robustness. Existing attacks are usually realized through word-level or sentence-level perturbations, which either limit the perturbation space or sacrifice fluency and textual quality, both affecting the attack effectiveness. In this paper, we propose Phrase-Level Textual Adversarial ATtack (PLAT) that generates adversarial samples through phrase-level perturbations. PLAT first extracts the vulnerable phrases as attack targets by a syntactic parser, and then perturbs them by a pre-trained blank-infilling model. Such flexible perturbation design substantially expands the search space for more effective attacks without introducing too many modifications, and meanwhile maintaining the textual fluency and grammaticality via contextualized generation using surrounding texts. Moreover, we develop a label preservation filter leveraging the likelihoods of language models fine-tuned on each class, rather than textual similarity, to rule out those perturbations that potentially alter the original class label for humans. Extensive experiments and human evaluation demonstrate that PLAT has a superior attack effectiveness as well as a better label consistency than strong baselines.

Specializing Pre-trained Language Models for Better Relational Reasoning via Network Pruning
Siyu Ren and Kenny Q. Zhu 08:00-09:00 (702 Clearwater)
Pretrained masked language models (PLMs) were shown to be inheriting a considerable amount of relational knowledge from the source corpora. In this paper, we present an in-depth and comprehensive study concerning specializing PLMs into relational models from the perspective of network pruning. We show that it is possible to find subnetworks capable of representing grounded commonsense relations at non-trivial sparsity while being more generalizable than original PLMs in scenarios requiring knowledge of single or multiple commonsense relations.

Abstraction not Memory: BERT and the English Article System
Harish Tayyar Madabushi, Dagmar Divjak and Petar Milin 08:00-09:00 (702 Clearwater)
Article prediction is a task that has long defied accurate linguistic description. As such, this task is ideally suited to evaluate models on their ability to emulate native-speaker intuition. To this end, we compare the performance of native English speakers and pre-trained models on the task of article prediction set up as a three way choice (a/an, the, zero). Our experiments with BERT show that BERT outperforms humans on this task across all articles. In particular, BERT is far superior to humans at detecting the zero article, possibly because we insert them using rules that the deep neural model can easily pick up. More interestingly, we find that BERT tends to agree more with annotators than with the corpus when inter-annotator agreement is high but switches to agreeing more with the corpus as inter-annotator agreement drops. We contend that this alignment with annotators, despite being trained on the corpus, suggests that BERT is not memorising article use, but captures a high level generalisation of article use akin to human intuition.

Analysing the Correlation between Lexical Ambiguity and Translation Quality in a Multimodal Setting using WordNet
Ali Hatami, Paul Buitelaar and Mihael Arcan 08:00-09:00 (702 Clearwater)
Multimodal Neural Machine Translation is focusing on using visual information to translate sentences in the source language into the target

language. The main idea is to utilise information from visual modalities to promote the output quality of the text-based translation model. Although the recent multimodal strategies extract the most relevant visual information in images, the effectiveness of using visual information on translation quality changes based on the text dataset. Due to this, this work studies the impact of leveraging visual information in multimodal translation models of ambiguous sentences. Our experiments analyse the Multi30k evaluation dataset and calculate ambiguity scores of sentences based on the WordNet hierarchical structure. To calculate the ambiguity of a sentence, we extract the ambiguity scores for all nouns based on the number of senses in WordNet. The main goal is to find in which sentences, visual content can improve the text-based translation model. We report the correlation between the ambiguity scores and translation quality extracted for all sentences in the English-German dataset.

Language Model Augmented Monotonic Attention for Simultaneous Translation
Sathish Reddy Indurthi, Mohd Abbas Zaidi, Beomseok Lee, Nikhil Kumar Lakumarapu and Sangha Kim 08:00-09:00 (702 Clearwater)
The state-of-the-art adaptive policies for Simultaneous Neural Machine Translation (SNMT) use monotonic attention to perform read/write decisions based on the partial source and target sequences. The lack of sufficient information might cause the monotonic attention to take poor read/write decisions, which in turn negatively affects the performance of the SNMT model. On the other hand, human translators make better read/write decisions since they can anticipate the immediate future words using linguistic information and domain knowledge. In this work, we propose a framework to aid monotonic attention with an external language model to improve its decisions. Experiments on MuST-C English-German and English-French speech-to-text translation tasks show the future information from the language model improves the state-of-the-art monotonic multi-head attention model further.

Aligning Generative Language Models with Human Values
Ruibo Liu, Ge Zhang, Xinyu Feng and Soroush Vosoughi 08:00-09:00 (702 Clearwater)
Although current large-scale generative language models (LMs) can show impressive insights about factual knowledge, they do not exhibit similar success with respect to human values judgements (e.g., whether or not the generations of an LM are moral). Existing methods learn human values either by directly mimicking the behavior of human data, or rigidly constraining the generation space to human-chosen tokens. These methods are inherently limited in that they do not consider the contextual and abstract nature of human values and as a result often fail when dealing with out-of-domain context or sophisticated and abstract human values.

This paper proposes SENSEI, a new reinforcement learning based method that can embed human values judgements into each step of language generation. SENSEI deploys an Actor-Critic framework, where the Critic is a reward distributor that simulates the reward assignment procedure of humans, while the Actor guides the generation towards the maximum reward direction. Compared with five existing methods in three human values alignment datasets, SENSEI not only achieves higher alignment performance in terms of both automatic and human evaluations, but also shows improvements on robustness and transfer learning on unseen human values.

Non-Autoregressive Neural Machine Translation with Consistency Regularization Optimized Variational Framework
Minghao Zhu, Junli Wang and Chungang Yan 08:00-09:00 (702 Clearwater)
Variational Autoencoder (VAE) is an effective framework to model the interdependency for non-autoregressive neural machine translation (NAT). One of the prominent VAE-based NAT frameworks, LaNMT, achieves great improvements to vanilla models, but still suffers from two main issues which lower down the translation quality: (1) mismatch between training and inference circumstances and (2) inadequacy of latent representations. In this work, we target on addressing these issues by proposing posterior consistency regularization. Specifically, we first perform stochastic data augmentation on the input samples to better adapt the model for inference circumstance, and then conduct consistency training on posterior latent variables to construct a more robust latent representations without any expansion on latent size. Experiments on En<->De and En<->Ro benchmarks confirm the effectiveness of our methods with about 1.5/0.7 and 0.8/0.3 BLEU points improvement to the baseline model with about $12.6\times$ faster than autoregressive Transformer.

Cheat Codes to Quantify Missing Source Information in Neural Machine Translation
Proyag Pal and Kenneth Heafield 08:00-09:00 (702 Clearwater)
This paper describes a method to quantify the amount of information $H(t|s)$ added by the target sentence t that is not present in the source s in a neural machine translation system. We do this by providing the model the target sentence in a highly compressed form (a "cheat code"), and exploring the effect of the size of the cheat code. We find that the model is able to capture extra information from just a single float representation of the target and nearly reproduces the target with two 32-bit floats per target token.

Training Mixed-Domain Translation Models via Federated Learning
Peyman Passban, Tanya Roosta, Rahul Gupta, Ankit Chadha and Clement Chung 08:00-09:00 (702 Clearwater)
Training mixed-domain translation models is a complex task that demands tailored architec- tures and costly data preparation techniques. In this work, we leverage federated learning (FL) in order to tackle the problem. Our investiga- tion demonstrates that with slight modifications in the training process, neural machine trans- lation (NMT) engines can be easily adapted when an FL-based aggregation is applied to fuse different domains. Experimental results also show that engines built via FL are able to perform on par with state-of-the-art baselines that rely on centralized training techniques. We evaluate our hypothesis in the presence of five datasets with different sizes, from different domains, to translate from German into English and discuss how FL and NMT can mutually benefit from each other. In addition to provid- ing benchmarking results on the union of FL and NMT, we also propose a novel technique to dynamically control the communication band- width by selecting impactful parameters during FL updates. This is a significant achievement considering the large size of NMT engines that need to be exchanged between FL parties.

Bi-SimCut: A Simple Strategy for Boosting Neural Machine Translation
Pengzhi Gao, Zhongjun He, Hua Wu and Haifeng Wang 08:00-09:00 (702 Clearwater)
We introduce Bi-SimCut: a simple but effective training strategy to boost neural machine translation (NMT) performance. It consists of two procedures: bidirectional pretraining and unidirectional finetuning. Both procedures utilize SimCut, a simple regularization method that forces the consistency between the output distributions of the original and the cutoff sentence pairs. Without leveraging extra dataset via back-translation or integrating large-scale pretrained model, Bi-SimCut achieves strong translation performance across five translation benchmarks (data sizes range from 160K to 20.2M): BLEU scores of 31.16 for $\texttt{en} \rightarrow \texttt{de}$ and 38.37 for $\texttt{de} \rightarrow \texttt{en}$ on the IWSLT14 dataset, 30.78 for $\texttt{en} \rightarrow \texttt{de}$ and 35.15 for $\texttt{de} \rightarrow \texttt{en}$ on the WMT14 dataset, and 27.17 for $\texttt{zh} \rightarrow \texttt{en}$ on the WMT17 dataset. SimCut is not a new method, but a version of Cutoff (Shen et al., 2020) simplified and adapted for NMT, and it could be considered as a perturbation-based method. Given the universality and simplicity of Bi-SimCut and SimCut, we believe they can serve as strong baselines for future NMT research.

Latent Group Dropout for Multilingual and Multidomain Machine Translation
Minh-Quang Pham, François Yvon and Josep Crego 08:00-09:00 (702 Clearwater)
Multidomain and multilingual machine translation often rely on parameter sharing strategies, where large portions of the network are meant to capture the commonalities of the tasks at hand, while smaller parts are reserved to model the peculiarities of a language or a domain. In adapter-based approaches, these strategies are hardcoded in the network architecture, independent of the similarities between tasks. In this work, we propose a new method to better take advantage of these similarities, using a latent-variable model. We also develop new techniques

to train this model end-to-end and report experimental results showing that the learned patterns are both meaningful and yield improved translation performance without any increase of the model size.

Bridging the Gap between Training and Inference: Multi-Candidate Optimization for Diverse Neural Machine Translation
Huan Lin, Baosong Yang, Liang Yao, Dayiheng Liu, Haibo Zhang, Jun Xie, Min Zhang and Jinsong Su 08:00-09:00 (702 Clearwater)
Diverse NMT aims at generating multiple diverse yet faithful translations given a source sentence. In this paper, we investigate a common shortcoming in existing diverse NMT studies: the model is usually trained with single reference, while expected to generate multiple candidate translations in inference. The discrepancy between training and inference enlarges the confidence variance and quality gap among candidate translations and thus hinders model performance. To deal with this defect, we propose a multi-candidate optimization framework for diverse NMT. Specifically, we define assessments to score the diversity and the quality of candidate translations during training, and optimize the diverse NMT model with two strategies based on reinforcement learning, namely hard constrained training and soft constrained training. We conduct experiments on NIST Chinese-English and WMT14 English-German translation tasks. The results illustrate that our framework is transparent to basic diverse NMT models, and universally makes better trade-off between diversity and quality. Our source codeis available at https://github.com/DeepLearnXMU/MultiCanOptim.

Inducing and Using Alignments for Transition-based AMR Parsing
Andrew Drozdov, Jiawei Zhou, Radu Florian, Andrew McCallum, Tahira Naseem, Yoon Kim and Ramón Fernandez Astudillo 08:00-09:00 (702 Clearwater)
Transition-based parsers for Abstract Meaning Representation (AMR) rely on node-to-word alignments. These alignments are learned separately from parser training and require a complex pipeline of rule-based components, pre-processing, and post-processing to satisfy domain-specific constraints. Parsers also train on a point-estimate of the alignment pipeline, neglecting the uncertainty due to the inherent ambiguity of alignment. In this work we explore two avenues for overcoming these limitations. First, we propose a neural aligner for AMR that learns node-to-word alignments without relying on complex pipelines. We subsequently explore a tighter integration of aligner and parser training by considering a distribution over oracle action sequences arising from aligner uncertainty. Empirical results show this approach leads to more accurate alignments and generalization better from the AMR2.0 to AMR3.0 corpora. We attain a new state-of-the art for gold-only trained models, matching silver-trained performance without the need for beam search on AMR3.0.

Label Anchored Contrastive Learning for Language Understanding
Zhenyu Zhang, Yuming Zhao, Meng Chen and Xiaodong He 08:00-09:00 (702 Clearwater)
Contrastive learning (CL) has achieved astonishing progress in computer vision, speech, and natural language processing fields recently with self-supervised learning. However, CL approach to the supervised setting is not fully explored, especially for the natural language understanding classification task. Intuitively, the class label itself has the intrinsic ability to perform hard positive/negative mining, which is crucial for CL. Motivated by this, we propose a novel label anchored contrastive learning approach (denoted as LaCon) for language understanding. Specifically, three contrastive objectives are devised, including a multi-head instance-centered contrastive loss (ICL), a label-centered contrastive loss (LCL), and a label embedding regularizer (LER). Our approach does not require any specialized network architecture or any extra data augmentation, thus it can be easily plugged into existing powerful pre-trained language models. Compared to the state-of-the-art baselines, LaCon obtains up to 4.1% improvement on the popular datasets of GLUE and CLUE benchmarks. Besides, LaCon also demonstrates significant advantages under the few-shot and data imbalance settings, which obtains up to 9.4% improvement on the FewGLUE and FewCLUE benchmarking tasks.

Improving Constituent Representation with Hypertree Neural Networks
Hao Zhou, Gongshen Liu and Kewei Tu 08:00-09:00 (702 Clearwater)
Many natural language processing tasks involve text spans and thus high-quality span representations are needed to enhance neural approaches to these tasks. Most existing methods of span representation are based on simple derivations (such as max-pooling) from word representations and do not utilize compositional structures of natural language. In this paper, we aim to improve representations of constituent spans using a novel hypertree neural networks (HTNN) that is structured with constituency parse trees. Each node in the HTNN represents a constituent of the input sentence and each hyperedge represents a composition of smaller child constituents into a larger parent constituent. In each update iteration of the HTNN, the representation of each constituent is computed based on all the hyperedges connected to it, thus incorporating both bottom-up and top-down compositional information. We conduct comprehensive experiments to evaluate HTNNs against other span representation models and the results show the effectiveness of HTNN.

Generic and Trend-aware Curriculum Learning for Relation Extraction
Nidhi Vakil and Hadi Amiri 08:00-09:00 (702 Clearwater)
We present a generic and trend-aware curriculum learning approach that effectively integrates textual and structural information in text graphs for relation extraction between entities, which we consider as node pairs in graphs. The proposed model extends existing curriculum learning approaches by incorporating sample-level loss trends to better discriminate easier from harder samples and schedule them for training. The model results in a robust estimation of sample difficulty and shows sizable improvement over the state-of-the-art approaches across several datasets.

On the Effectiveness of Sentence Encoding for Intent Detection Meta-Learning
Tingting Ma, Qianhui Wu, Zhiwei Yu, Tiejun Zhao and Chin-Yew Lin 08:00-09:00 (702 Clearwater)
Recent studies on few-shot intent detection have attempted to formulate the task as a meta-learning problem, where a meta-learning model is trained with a certain capability to quickly adapt to newly specified few-shot tasks with potentially unseen intent categories. Prototypical networks have been commonly used in this setting, with the hope that good prototypical representations could be learned to capture the semantic similarity between the query and a few labeled instances. This intuition naturally leaves a question of whether or not a good sentence representation scheme could suffice for the task without further domain-specific adaptation. In this paper, we conduct empirical studies on a number of general-purpose sentence embedding schemes, showing that good sentence embeddings without any fine-tuning on intent detection data could produce a non-trivially strong performance. Inspired by the results from our qualitative analysis, we propose a frustratingly easy modification, which leads to consistent improvements over all sentence encoding schemes, including those from the state-of-the-art prototypical network variants with task-specific fine-tuning.

A Data Cartography based MixUp for Pre-trained Language Models
Seo Yeon Park and Cornelia Caragea 08:00-09:00 (702 Clearwater)
MixUp is a data augmentation strategy where additional samples are generated during training by combining random pairs of training samples and their labels. However, selecting random pairs is not potentially an optimal choice. In this work, we propose TDMixUp, a novel MixUp strategy that leverages Training Dynamics and allows more informative samples to be combined for generating new data samples. Our proposed TDMixUp first measures confidence, variability, (Swayamdipta et al., 2020), and Area Under the Margin (AUM) (Pleiss et al., 2020) to identify the characteristics of training samples (e.g., as easy-to-learn or ambiguous samples), and then interpolates these characterized samples. We empirically validate that our method not only achieves competitive performance using a smaller subset of the training data

compared with strong baselines, but also yields lower expected calibration error on the pre-trained language model, BERT, on both in-domain and out-of-domain settings in a wide range of NLP tasks. We publicly release our code.

Embedding Hallucination for Few-shot Language Fine-tuning
Yiren Jian, Chongyang Gao and Soroush Vosoughi 08:00-09:00 (702 Clearwater)
Few-shot language learners adapt knowledge from a pre-trained model to recognize novel classes from a few-labeled sentences. In such settings, fine-tuning a pre-trained language model can cause severe over-fitting. In this paper, we propose an Embedding Hallucination (EmbedHalluc) method, which generates auxiliary embedding-label pairs to expand the fine-tuning dataset. The hallucinator is trained by playing an adversarial game with the discriminator, such that the hallucinated embedding is indiscriminative to the real ones in the fine-tuning dataset. By training with the extended dataset, the language learner effectively learns from the diverse hallucinated embeddings to overcome the over-fitting issue. Experiments demonstrate that our proposed method is effective in a wide range of language tasks, outperforming current fine-tuning methods. Further, we show that EmbedHalluc outperforms other methods that address this over-fitting problem, such as common data augmentation, semi-supervised pseudo-labeling, and regularization.

Contrastive Learning for Prompt-based Few-shot Language Learners
Yiren Jian, Chongyang Gao and Soroush Vosoughi 08:00-09:00 (702 Clearwater)
The impressive performance of GPT-3 using natural language prompts and in-context learning has inspired work on better fine-tuning of moderately-sized models under this paradigm. Following this line of work, we present a contrastive learning framework that clusters inputs from the same class for better generality of models trained with only limited examples. Specifically, we propose a supervised contrastive framework that clusters inputs from the same class under different augmented "views" and repel the ones from different classes. We create different "views" of an example by appending it with different language prompts and contextual demonstrations. Combining a contrastive loss with the standard masked language modeling (MLM) loss in prompt-based few-shot learners, the experimental results show that our method can improve over the state-of-the-art methods in a diverse set of 15 language tasks. Our framework makes minimal assumptions on the task or the base model, and can be applied to many recent methods with little modification.

Consistency Training with Virtual Adversarial Discrete Perturbation
Jungsoo Park, Gyuwan Kim and Jaewoo Kang 08:00-09:00 (702 Clearwater)
Consistency training regularizes a model by enforcing predictions of original and perturbed inputs to be similar. Previous studies have proposed various augmentation methods for the perturbation but are limited in that they are agnostic to the training model. Thus, the perturbed samples may not aid in regularization due to their ease of classification from the model. In this context, we propose an augmentation method of adding a discrete noise that would incur the highest divergence between predictions. This virtual adversarial discrete noise obtained by replacing a small portion of tokens while keeping original semantics as much as possible efficiently pushes a training model's decision boundary. Experimental results show that our proposed method outperforms other consistency training baselines with text editing, paraphrasing, or a continuous noise on semi-supervised text classification tasks and a robustness benchmark.

Embarrassingly Simple Performance Prediction for Abductive Natural Language Inference
Emīls Kadiķis, Vaibhav Srivastav and Roman Klinger 08:00-09:00 (702 Clearwater)
The task of natural language inference (NLI), to decide if a hypothesis entails or contradicts a premise, received considerable attention in recent years. All competitive systems build on top of contextualized representations and make use of transformer architectures for learning an NLI model. When somebody is faced with a particular NLI task, they need to select the best model that is available. This is a time-consuming and resource-intense endeavour. To solve this practical problem, we propose a simple method for predicting the performance without actually fine-tuning the model. We do this by testing how well the pre-trained models perform on the aNLI task when just comparing sentence embeddings with cosine similarity to what kind of performance is achieved when training a classifier on top of these embeddings. We show that the accuracy of the cosine similarity approach correlates strongly with the accuracy of the classification approach with a Pearson correlation coefficient of 0.65. Since the similarity is orders of magnitude faster to compute on a given dataset (less than a minute vs. hours), our method can lead to significant time savings in the process of model selection.

Regularized Training of Nearest Neighbor Language Models
Jean-Francois Ton, Walter Talbott, Shuangfei Zhai and Joshua M. Susskind 08:00-09:00 (702 Clearwater)
Including memory banks in a natural language processing architecture increases model capacity by equipping it with additional data at inference time. In this paper, we build upon kNN-LM, which uses a pre-trained language model together with an exhaustive kNN search through the training data (memory bank) to achieve state-of-the-art results. We investigate whether we can improve the kNN-LM performance by instead training a LM with the knowledge that we will be using a kNN post-hoc. We achieved significant improvement using our method on language modeling tasks on extttWIKI-2 and extttWIKI-103. The main phenomenon that we encounter is that adding a simple L2 regularization on the activations (not weights) of the model, a transformer, improves the post-hoc kNN classification performance. We explore some possible reasons for this improvement. In particular, we find that the added L2 regularization seems to improve the performance for high-frequency words without deteriorating the performance for low frequency ones.

On Curriculum Learning for Commonsense Reasoning
Adyasha Maharana and Mohit Bansal 08:00-09:00 (702 Clearwater)
Commonsense reasoning tasks follow a standard paradigm of finetuning pretrained language models on the target task data, where samples are introduced to the model in a random order during training. However, recent research suggests that data order can have a significant impact on the performance of finetuned models for natural language understanding. Hence, we examine the effect of a human-like easy-to-difficult curriculum during finetuning of language models for commonsense reasoning tasks. We use paced curriculum learning to rank data and sample training mini-batches with increasing levels of difficulty from the ranked dataset during finetuning. Further, we investigate the effect of an adaptive curriculum, i.e., the data ranking is dynamically updated during training based on the current state of the learner model. We use a teacher model to measure difficulty of each sample and experiment with three measures based on question answering probability, variability and out-of-distribution. To understand the effectiveness of curriculum learning in various scenarios, we apply it on full model fine-tuning as well as parameter-efficient prompt-tuning settings. Our results show that fixed as well as adaptive curriculum learning significantly improve performance for five commonsense reasoning tasks, i.e., SocialIQA, CosmosQA, CODAH, HellaSwag, WinoGrande in both tuning settings. Further, we find that prioritizing the difficult samples in the tail end of training improves generalization to unseen in-domain data as well as out-of-domain data. Our work provides evidence and encourages research into curriculum learning for commonsense reasoning.

Efficient Hierarchical Domain Adaptation for Pretrained Language Models
Alexandra Chronopoulou, Matthew E Peters and Jesse Dodge 08:00-09:00 (702 Clearwater)
The remarkable success of large language models has been driven by dense models trained on massive unlabeled, unstructured corpora. These corpora typically contain text from diverse, heterogeneous sources, but information about the source of the text is rarely used during training. Transferring their knowledge to a target domain is typically done by continuing training in-domain. In this paper, we introduce a method to permit domain adaptation to many diverse domains using a computationally efficient adapter approach. Our method is based on the

observation that textual domains are partially overlapping, and we represent domains as a hierarchical tree structure where each node in the tree is associated with a set of adapter weights. When combined with a frozen pretrained language model, this approach enables parameter sharing among related domains, while avoiding negative interference between unrelated ones. Experimental results with GPT-2 and a large fraction of the 100 most represented websites in C4 show across-the-board improvements in-domain. We additionally provide an inference time algorithm for a held-out domain and show that averaging over multiple paths through the tree enables further gains in generalization, while adding only a marginal cost to inference.

Improving In-Context Few-Shot Learning via Self-Supervised Training
Mingda Chen, Jingfei Du, Ramakanth Pasunuru, Todor Mihaylov, Srini Iyer, Veselin Stoyanov and Zornitsa Kozareva 08:00-09:00 (702 Clearwater)
Self-supervised pretraining has made few-shot learning possible for many NLP tasks. But the pretraining objectives are not typically adapted specifically for in-context few-shot learning. In this paper, we propose to use self-supervision in an intermediate training stage between pretraining and downstream few-shot usage with the goal to teach the model to perform in-context few shot learning. We propose and evaluate four self-supervised objectives on two benchmarks. We find that the intermediate self-supervision stage produces models that outperform strong baselines. Ablation study shows that several factors affect the downstream performance, such as the amount of training data and the diversity of the self-supervised objectives. Human-annotated cross-task supervision and self-supervision are complementary. Qualitative analysis suggests that the self-supervised-trained models are better at following task requirements.

Learning to Generate Examples for Semantic Processing Tasks
Danilo Croce, Simone Filice, Giuseppe Castellucci and Roberto Basili 08:00-09:00 (702 Clearwater)
Even if recent Transformer-based architectures, such as BERT, achieved impressive results in semantic processing tasks, their fine-tuning stage still requires large scale training resources. Usually, Data Augmentation (DA) techniques can help to deal with low resource settings. In Text Classification tasks, the objective of DA is the generation of well-formed sentences that i) represent the desired task category and ii) are novel with respect to existing sentences. In this paper, we propose a neural approach to automatically learn to generate new examples using a pre-trained sequence-to-sequence model. We first learn a task-oriented similarity function that we use to pair similar examples. Then, we use these example pairs to train a model to generate examples. Experiments in low resource settings show that augmenting the training material with the proposed strategy systematically improves the results on text classification and natural language inference tasks by up to 10% accuracy, outperforming existing DA approaches.

On the Effect of Pretraining Corpora on In-context Learning by a Large-scale Language Model
Seongjin Shin, Sang-Woo Lee, Hwijeen Ahn, Sungdong Kim, HyoungSeok Kim, Boseop Kim, Kyunghyun Cho, Gichang Lee, Woomyoung Park, Jung-Woo Ha and Nako Sung 08:00-09:00 (702 Clearwater)
Many recent studies on large-scale language models have reported successful in-context zero- and few-shot learning ability. However, the in-depth analysis of when in-context learning occurs is still lacking. For example, it is unknown how in-context learning performance changes as the training corpus varies. Here, we investigate the effects of the source and size of the pretraining corpus on in-context learning in HyperCLOVA, a Korean-centric GPT-3 model. From our in-depth investigation, we introduce the following observations: (1) in-context learning performance heavily depends on the corpus domain source, and the size of the pretraining corpus does not necessarily determine the emergence of in-context learning, (2) in-context learning ability can emerge when a language model is trained on a combination of multiple corpora, even when each corpus does not result in in-context learning on its own, (3) pretraining with a corpus related to a downstream task does not always guarantee the competitive in-context learning performance of the downstream task, especially in the few-shot setting, and (4) the relationship between language modeling (measured in perplexity) and in-context learning does not always correlate: e.g., low perplexity does not always imply high in-context few-shot learning performance.

Zero-shot Cross-lingual Conversational Semantic Role Labeling
Han Wu, Haochen Tan, Kun Xu, Shuqi Liu, Lianwei Wu and Linqi Song 08:00-09:00 (702 Clearwater)
While conversational semantic role labeling (CSRL) has shown its usefulness on Chinese conversational tasks, it is still under-explored in non-Chinese languages due to the lack of multilingual CSRL annotations for the parser training. To avoid expensive data collection and error-propagation of translation-based methods, we present a simple but effective approach to perform zero-shot cross-lingual CSRL. Our model implicitly learns language-agnostic, conversational structure-aware and semantically rich representations with the hierarchical encoders and elaborately designed pre-training objectives. Experimental results show that our model outperforms all baselines by large margins on two newly collected English CSRL test sets. More importantly, we confirm the usefulness of CSRL to non-Chinese conversational tasks such as the question-in-context rewriting task in English and the multi-turn dialogue response generation tasks in English, German and Japanese by incorporating the CSRL information into the downstream conversation-based models. We believe this finding is significant and will facilitate the research of non-Chinese dialogue tasks which suffer the problems of ellipsis and anaphora.

Explaining Why: How Instructions and User Interfaces Impact Annotator Rationales When Labeling Text Data
Cynthia Sullivan, William Brackenbury, Andrew Michael McNut, Kevin Bryson, kbyllofficial@gmail.com kbyllofficial@gmail.com, Yuxin Chen, Michael Littman, Chenhao Tan and Blase Ur 08:00-09:00 (702 Clearwater)
In the context of data labeling, NLP researchers are increasingly interested in having humans select rationales, a subset of input tokens relevant to the chosen label. We conducted a 332-participant online user study to understand how humans select rationales, especially how different instructions and user interface affordances impact the rationales chosen. Participants labeled ten movie reviews as positive or negative, selecting words and phrases supporting their label as rationales. We varied the instructions given, the rationale-selection task, and the user interface. Participants often selected about 12% of input tokens as rationales, but selected fewer if unable to drag over multiple tokens at once. Whereas participants were near unanimous in their data labels, they were far less consistent in their rationales. The user interface affordances and task greatly impacted the types of rationales chosen. We also observed large variance across participants.

Anti-Overestimation Dialogue Policy Learning for Task-Completion Dialogue System
Chang Tian, Wenpeng Yin and Marie-Francine Moens 08:00-09:00 (702 Clearwater)
A dialogue policy module is an essential part of task-completion dialogue systems. Recently, increasing interest has focused on reinforcement learning (RL)-based dialogue policy. Its favorable performance and wise action decisions rely on an accurate estimation of action values. The overestimation problem is a widely known issue of RL since its estimate of the maximum action value is larger than the ground truth, which results in an unstable learning process and suboptimal policy. This problem is detrimental to RL-based dialogue policy learning. To mitigate this problem, this paper proposes a dynamic partial average estimator (DPAV) of the ground truth maximum action value. DPAV calculates the partial average between the predicted maximum action value and minimum action value, where the weights are dynamically adaptive and problem-dependent. We incorporate DPAV into a deep Q-network as the dialogue policy and show that our method can achieve better or comparable results compared to top baselines on three dialogue datasets of different domains with a lower computational load. In addition, we also theoretically prove the convergence and derive the upper and lower bounds of the bias compared with those of other methods.

Prompt Augmented Generative Replay via Supervised Contrastive Learning for Lifelong Intent Detection

Vaibhav Varshney, Mayur Patidar, Rajat Kumar, Lovekesh Vig and Gautam Shroff 08:00-09:00 (702 Clearwater)
Identifying all possible user intents for a dialog system at design time is challenging even for skilled domain experts. For practical applications, novel intents may have to be inferred incrementally on the fly. This typically entails repeated retraining of the intent detector on both the existing and novel intents which can be expensive and would require storage of all past data corresponding to prior intents. In this paper, the objective is to continually train an intent detector on new intents while maintaining performance on prior intents without mandating access to prior intent data. Several data replay-based approaches have been introduced to avoid catastrophic forgetting during continual learning, including exemplar and generative replay. Current generative replay approaches struggle to generate representative samples because the generation is conditioned solely on the class/task label. Motivated by the recent work around prompt-based generation via pre-trained language models (PLMs), we employ generative replay using PLMs for incremental intent detection. Unlike exemplar replay, we only store the relevant contexts per intent in memory and use these stored contexts (with the class label) as prompts for generating intent-specific utterances. We use a common model for both generation and classification to promote optimal sharing of knowledge across both tasks. To further improve generation, we employ supervised contrastive fine-tuning of the PLM. Our proposed approach achieves state-of-the-art (SOTA) for lifelong intent detection on four public datasets and even outperforms exemplar replay-based approaches. The technique also achieves SOTA on a lifelong relation extraction task, suggesting that the approach is extendable to other continual learning tasks beyond intent detection.

NLU++: A Multi-Label, Slot-Rich, Generalisable Dataset for Natural Language Understanding in Task-Oriented Dialogue
Inigo Casanueva, Ivan Vulić, Georgios P. Spithourakis and Paweł Budzianowski 08:00-09:00 (702 Clearwater)
We present NLU++, a novel dataset for natural language understanding (NLU) in task-oriented dialogue (ToD) systems, with the aim to provide a much more challenging evaluation environment for dialogue NLU models, up to date with the current application and industry requirements. NLU++ is divided into two domains (BANKING and HOTELS) and brings several crucial improvements over current commonly used NLU datasets. 1) NLU++ provides fine-grained domain ontologies with a large set of challenging multi-intent sentences combined with finer-grained and thus more challenging slot sets. 2) The ontology is divided into domain-specific and generic (i.e., domain-universal) intents that overlap across domains, promoting cross-domain reusability of annotated examples. 3) The dataset design has been inspired by the problems observed in industrial ToD systems, and 4) it has been collected, filtered and carefully annotated by dialogue NLU experts, yielding high-quality annotated data. Finally, we benchmark a series of current state-of-the-art NLU models on NLU++; the results demonstrate the challenging nature of the dataset, especially in low-data regimes, and call for further research on ToD NLU.

SKILL: Structured Knowledge Infusion for Large Language Models
Fedor Moiseev, Zhe Dong, Enrique Alfonseca and Martin Jaggi 08:00-09:00 (702 Clearwater)
Large language models (LLMs) have demonstrated human-level performance on a vast spectrum of natural language tasks. However, it is largely unexplored whether they can better internalize knowledge from a structured data, such as a knowledge graph, or from text. In this work, we propose a method to infuse structured knowledge into LLMs, by directly training T5 models on factual triples of knowledge graphs (KGs). We show that models pre-trained on Wikidata KG with our method outperform the T5 baselines on FreebaseQA and WikiHop, as well as the Wikidata-answerable subset of TriviaQA and NaturalQuestions. The models pre-trained on factual triples compare competitively with the ones on natural language sentences that contain the same knowledge. Trained on a smaller size KG, WikiMovies, we saw 3x improvement of exact match score on MetaQA task. The proposed method has an advantage that no alignment between the knowledge graph and text corpus is required in curating training data. This makes our method particularly useful when working with industry-scale knowledge graphs.

Collective Relevance Labeling for Passage Retrieval
Jihyuk Kim, Minsoo Kim and Seung-won Hwang 08:00-09:00 (702 Clearwater)
Deep learning for Information Retrieval (IR) requires a large amount of high-quality query-document relevance labels, but such labels are inherently sparse. Label smoothing redistributes some observed probability mass over unobserved instances, often uniformly, uninformed of the true distribution. In contrast, we propose knowledge distillation for informed labeling, without incurring high computation overheads at evaluation time. Our contribution is designing a simple but efficient teacher model which utilizes collective knowledge, to outperform state-of-the-arts distilled from a more complex teacher model. Specifically, we train up to $\times 8$ faster than the state-of-the-art teacher, while distilling the rankings better. Our code is publicly available at `https://github.com/jihyukkim-nlp/CollectiveKD`.

Domain-matched Pre-training Tasks for Dense Retrieval
Barlas Oguz, Kushal Lakhotia, Anchit Gupta, Patrick Lewis, Vladimir Karpukhin, Aleksandra Piktus, Xilun Chen, Sebastian Riedel, Scott Yih, Sonal Gupta and Yashar Mehdad 08:00-09:00 (702 Clearwater)
Pre-training on larger datasets with ever increasing model size is now a proven recipe for increased performance across almost all NLP tasks. A notable exception is information retrieval, where additional pre-training has so far failed to produce convincing results. We show that, with the right pre-training setup, this barrier can be overcome. We demonstrate this by pre-training large bi-encoder models on 1) a recently released set of 65 million synthetically generated questions, and 2) 200 million post-comment pairs from a preexisting dataset of Reddit conversations made available by pushshift.io. We evaluate on a set of information retrieval and dialogue retrieval benchmarks, showing substantial improvements over supervised baselines.

CL-ReLKT: Cross-lingual Language Knowledge Transfer for Multilingual Retrieval Question Answering
Peerat Limkonchotiwat, Wuttikorn Ponwitayarat, Can Udomcharoenchaikit, Ekapol Chuangsuwanich and Sarana Nutanong 08:00-09:00 (702 Clearwater)
Cross-Lingual Retrieval Question Answering (CL-ReQA) is concerned with retrieving answer documents or passages to a question written in a different language. A common approach to CL-ReQA is to create a multilingual sentence embedding space such that question-answer pairs across different languages are close to each other. In this paper, we propose a novel CL-ReQA method utilizing the concept of language knowledge transfer and a new cross-lingual consistency training technique to create a multilingual embedding space for ReQA. To assess the effectiveness of our work, we conducted comprehensive experiments on CL-ReQA and a downstream task, machine reading QA. We compared our proposed method with the current state-of-the-art solutions across three public CL-ReQA corpora. Our method outperforms competitors in 19 out of 21 settings of CL-ReQA. When used with a downstream machine reading QA task, our method outperforms the best existing language-model-based method by 10% in F1 while being 10 times faster in sentence embedding computation. The code and models are available at https://github.com/mrpeerat/CL-ReLKT.

Weakly Supervised Text Classification using Supervision Signals from a Language Model
Ziqian Zeng, Weimin Ni, Tianqing Fang, Xiang Li, Xinran Zhao and Yangqiu Song 08:00-09:00 (702 Clearwater)
Solving text classification in a weakly supervised manner is important for real-world applications where human annotations are scarce. In this paper, we propose to query a masked language model with cloze style prompts to obtain supervision signals. We design a prompt which combines the document itself and "this article is talking about `[MASK]`." A masked language model can generate words for the `[MASK]` token. The generated words which summarize the content of a document can be utilized as supervision signals. We propose a latent variable model to learn a word distribution learner which associates generated words to pre-defined categories and a document classifier simultaneously without using any annotated data. Evaluation on three datasets, AGNews, 20Newsgroups, and UCINews, shows that our method can outperform baselines by 2%, 4%, and 3%.

A Generative Language Model for Few-shot Aspect-Based Sentiment Analysis
Ehsan Hosseini-Asl, Wenhao Liu and Caiming Xiong 08:00-09:00 (702 Clearwater)
Sentiment analysis is an important task in natural language processing. In recent works, pre-trained language models are often used to achieve state-of-the-art results, especially when training data is scarce. It is common to fine-tune on the downstream task, usually by adding task-specific layers on top of the model. In this paper, we focus on aspect-based sentiment analysis, which involves extracting aspect term, category, and predicting their corresponding polarities. In particular, we are interested in few-shot settings. We propose to reformulate the extraction and prediction tasks into the sequence generation task, using a generative language model with unidirectional attention (GPT2 is used unless stated otherwise). This way, the model learns to accomplish the tasks via language generation without the need of training task-specific layers. Our evaluation results on the single-task polarity prediction show that our approach outperforms the previous state-of-the-art (based on BERT) on average performance by a large margins in few-shot and full-shot settings. More importantly, our generative approach significantly reduces the model variance caused by low-resource data. We further demonstrate that the proposed generative language model can handle joint and multi-task settings, unlike previous work. We observe that the proposed sequence generation method achieves further improved performances on polarity prediction when the model is trained via joint and multi-task settings. Further evaluation on similar sentiment analysis datasets, SST-2, SST-5 and OOS intent detection validates the superiority and noise robustness of generative language model in few-shot settings.

RGL: A Simple yet Effective Relation Graph Augmented Prompt-based Tuning Approach for Few-Shot Learning
Yaqing Wang, Xin Tian, Haoyi Xiong, Yueyang Li, Zeyu Chen, Sheng Guo and Dejing Dou 08:00-09:00 (702 Clearwater)
Pre-trained language models (PLMs) can provide a good starting point for downstream applications. However, it is difficult to generalize PLMs to new tasks given a few labeled samples. In this work, we show that Relation Graph augmented Learning (RGL) can improve the performance of few-shot natural language understanding tasks. During learning, RGL constructs a relation graph based on the label consistency between samples in the same batch, and learns to solve the resultant node classification and link prediction problems on the relation graph. In this way, RGL fully exploits the limited supervised information, which can boost the tuning effectiveness. Extensive experimental results show that RGL consistently improves the performance of prompt-based tuning strategies.

Speeding Up Entmax
Maxat Tezekbayev, Vassilina Nikoulina, Matthias Gallé and Zhenisbek Assylbekov 08:00-09:00 (702 Clearwater)
Softmax is the de facto standard for normalizing logits in modern neural networks for language processing. However, by producing a dense probability distribution each token in the vocabulary has a nonzero chance of being selected at each generation step, leading to a variety of reported problems in text generation. α-entmax of Peters et al. (2019) solves this problem, but is unfortunately slower than softmax. In this paper, we propose an alternative to α-entmax, which keeps its virtuous characteristics, but is as fast as optimized softmax and achieves on par or better performance in machine translation task.

MixQG: Neural Question Generation with Mixed Answer Types
Lidiya Murakhovs'ka, Chien-Sheng Wu, Philippe Laban, Tong Niu, Wenhao Liu and Caiming Xiong 08:00-09:00 (702 Clearwater)
Asking good questions is an essential ability for both human and machine intelligence. However, existing neural question generation approaches mainly focus on short factoid type of answers. In this paper, we introduce a neural question generator, MixQG, to bridge this gap. We combine nine question answering datasets with diverse answer types, including yes/no, multiple-choice, extractive, and abstractive answers, to train a single generative model. We show with empirical results that our model outperforms existing work in both seen and unseen domains, and can generate questions with different cognitive levels when conditioned on different answer types. We run a human evaluation study to assess the quality of generated questions and find that MixQG outperforms the next best model by 10%. Our code and model checkpoints will be released and integrated with the HuggingFace library to facilitate various downstream applications.

SeaD: End-to-end Text-to-SQL Generation with Schema-aware Denoising
Kuan Xu, Yongbo Wang, Yongliang Wang, Zihao Wang, Zujie Wen and Yang Dong 08:00-09:00 (702 Clearwater)
On the WikiSQL benchmark, most methods tackle the challenge of text-to-SQL with predefined sketch slots and build sophisticated sub-tasks to fill these slots. Though achieving promising results, these methods suffer from over-complex model structure. In this paper, we present a simple yet effective approach that enables auto-regressive sequence-to-sequence model to robust text-to-SQL generation. Instead of formulating the task of text-to-SQL as slot-filling, we propose to train sequence-to-sequence model with Schema-aware Denoising (SeaD), which consists of two denoising objectives that train model to either recover input or predict output from two novel erosion and shuffle noises. These model-agnostic denoising objectives act as the auxiliary tasks for structural data modeling during sequence-to-sequence generation. In addition, we propose a clause-sensitive execution guided (EG) decoding strategy to overcome the limitation of EG decoding for generative model. The experiments show that the proposed method improves the performance of sequence-to-sequence model in both schema linking and grammar correctness and establishes new state-of-the-art on WikiSQL benchmark. Our work indicates that the capacity of sequence-to-sequence model for text-to-SQL may have been under-estimated and could be enhanced by specialized denoising task.

DecBERT: Enhancing the Language Understanding of BERT with Causal Attention Masks
Ziyang Luo, Yadong Xi, Jing Ma, Zhiwei Yang, Xiaoxi Mao, Changjie Fan and Rongsheng Zhang 08:00-09:00 (702 Clearwater)
Since 2017, the Transformer-based models play critical roles in various downstream Natural Language Processing tasks. However, a common limitation of the attention mechanism utilized in Transformer Encoder is that it cannot automatically capture the information of word order, so explicit position embeddings are generally required to be fed into the target model. In contrast, Transformer Decoder with the causal attention masks is naturally sensitive to the word order. In this work, we focus on improving the position encoding ability of BERT with the causal attention masks. Furthermore, we propose a new pre-trained language model *DecBERT* and evaluate it on the GLUE benchmark. Experimental results show that (1) the causal attention mask is effective for BERT on the language understanding tasks; (2) our *DecBERT* model without position embeddings achieve comparable performance on the GLUE benchmark; and (3) our modification accelerates the pre-training process and *DecBERT w/ PE* achieves better overall performance than the baseline systems when pre-training with the same amount of computational resources.

Session 8 - 09:15-10:15

Interpretability and Analysis of Models for NLP 3

09:15-10:15 (Columbia A)

Exploiting Inductive Bias in Transformers for Unsupervised Disentanglement of Syntax and Semantics with VAEs

Ghazi Felhi, Joseph Le Roux and Djamé Seddah 09:15-09:30 (Columbia A)

We propose a generative model for text generation, which exhibits disentangled latent representations of syntax and semantics. Contrary to previous work, this model does not need syntactic information such as constituency parses, or semantic information such as paraphrase pairs. Our model relies solely on the inductive bias found in attention-based architectures such as Transformers.
In the attention of Transformers, $keys$ handle information selection while $values$ specify what information is conveyed. Our model, dubbed QKVAE, uses Attention in its decoder to read latent variables where one latent variable infers keys while another infers values.
We run experiments on latent representations and experiments on syntax/semantics transfer which show that QKVAE displays clear signs of disentangled syntax and semantics. We also show that our model displays competitive syntax transfer capabilities when compared to supervised models and that comparable supervised models need a fairly large amount of data (more than 50K samples) to outperform it on both syntactic and semantic transfer. The code for our experiments is publicly available.

Time Waits for No One! Analysis and Challenges of Temporal Misalignment
Kelvin Luu, Daniel Khashabi, Suchin Gururangan, Karishma Mandyam and Noah Smith 09:30-09:45 (Columbia A)
When an NLP model is trained on text data from one time period and tested or deployed on data from another, the resulting temporal misalignment can degrade end-task performance. In this work, we establish a suite of eight diverse tasks across different domains (social media, science papers, news, and reviews) and periods of time (spanning five years or more) to quantify the effects of temporal misalignment. Our study is focused on the ubiquitous setting where a pretrained model is optionally adapted through continued domain-specific pretraining, followed by task-specific finetuning. We establish a suite of tasks across multiple domains to study temporal misalignment in modern NLP systems. We find stronger effects of temporal misalignment on task performance than have been previously reported. We also find that, while temporal adaptation through continued pretraining can help, these gains are small compared to task-specific finetuning on data from the target time period. Our findings motivate continued research to improve temporal robustness of NLP models.

What do Toothbrushes do in the Kitchen? How Transformers Think our World is Structured
Alexander Henlein and Alexander Mehler 09:45-10:00 (Columbia A)
Transformer-based models are now predominant in NLP. They outperform approaches based on static models in many respects. This success has in turn prompted research that reveals a number of biases in the language models generated by transformers. In this paper we utilize this research on biases to investigate to what extent transformer-based language models allow for extracting knowledge about object relations (X occurs in Y; X consists of Z; action A involves using X). To this end, we compare contextualized models with their static counterparts. We make this comparison dependent on the application of a number of similarity measures and classifiers. Our results are threefold: Firstly, we show that the models combined with the different similarity measures differ greatly in terms of the amount of knowledge they allow for extracting. Secondly, our results suggest that similarity measures perform much worse than classifier-based approaches. Thirdly, we show that, surprisingly, static models perform almost as well as contextualized models – in some cases even better.

A Study of the Attention Abnormality in Trojaned BERTs
Weimin Lyu, Songzhu Zheng, Tengfei Ma and Chao Chen 10:00-10:15 (Columbia A)
Trojan attacks raise serious security concerns. In this paper, we investigate the underlying mechanism of Trojaned BERT models. We observe the attention focus drifting behavior of Trojaned models, i.e., when encountering an poisoned input, the trigger token hijacks the attention focus regardless of the context. We provide a thorough qualitative and quantitative analysis of this phenomenon, revealing insights into the Trojan mechanism. Based on the observation, we propose an attention-based Trojan detector to distinguish Trojaned models from clean ones. To the best of our knowledge, we are the first to analyze the Trojan mechanism and develop a Trojan detector based on the transformer's attention.

Computational Social Science and Cultural Analytics

09:15-10:15 (Columbia C)

Mitigating Toxic Degeneration with Empathetic Data: Exploring the Relationship Between Toxicity and Empathy
Allison Lahnala, Charles Welch, Béla Neuendorf and Lucie Flek 09:15-09:30 (Columbia C)
Large pre-trained neural language models have supported the effectiveness of many NLP tasks, yet are still prone to generating toxic language hindering the safety of their use. Using empathetic data, we improve over recent work on controllable text generation that aims to reduce the toxicity of generated text. We find we are able to dramatically reduce the size of fine-tuning data to 7.5-30k samples while at the same time making significant improvements over state-of-the-art toxicity mitigation of up to 3.4% absolute reduction (26% relative) from the original work on 2.3m samples, by strategically sampling data based on empathy scores. We observe that the degree of improvements is subject to specific communication components of empathy. In particular, the more cognitive components of empathy significantly beat the original dataset in almost all experiments, while emotional empathy was tied to less improvement and even underperforming random samples of the original data. This is a particularly implicative insight for NLP work concerning empathy as until recently the research and resources built for it have exclusively considered empathy as an emotional concept.

Hatemoji: A Test Suite and Adversarially-Generated Dataset for Benchmarking and Detecting Emoji-Based Hate
Hannah Rose Kirk, Bertie Vidgen, Paul Rottger, Tristan Thrush and Scott A. Hale 09:30-09:45 (Columbia C)
Detecting online hate is a complex task, and low-performing models have harmful consequences when used for sensitive applications such as content moderation. Emoji-based hate is an emerging challenge for automated detection. We present HatemojiCheck, a test suite of 3,930 short-form statements that allows us to evaluate performance on hateful language expressed with emoji. Using the test suite, we expose weaknesses in existing hate detection models. To address these weaknesses, we create the HatemojiBuild dataset using a human-and-model-in-the-loop approach. Models built with these 5,912 adversarial examples perform substantially better at detecting emoji-based hate, while retaining strong performance on text-only hate. Both HatemojiCheck and HatemojiBuild are made publicly available.

A Holistic Framework for Analyzing the COVID-19 Vaccine Debate
Maria Leonor Pacheco, Tunazzina Islam, Monal Mahajan, Andrey Shor, Ming Yin, Lyle Ungar and Dan Goldwasser 09:45-10:00 (Columbia C)
The Covid-19 pandemic has led to infodemic of low quality information leading to poor health decisions. Combating the outcomes of this infodemic is not only a question of identifying false claims, but also reasoning about the decisions individuals make. In this work we propose a holistic analysis framework connecting stance and reason analysis, and fine-grained entity level moral sentiment analysis. We study how to model the dependencies between the different level of analysis and incorporate human insights into the learning process. Experiments show that our framework provides reliable predictions even in the low-supervision settings.

Hate Speech and Counter Speech Detection: Conversational Context Does Matter
Xinchen Yu, Eduardo Blanco and Lingzi Hong 10:00-10:15 (Columbia C)
Hate speech is plaguing the cyberspace along with user-generated content. Adding counter speech has become an effective way to combat hate speech online. Existing datasets and models target either (a) hate speech or (b) hate and counter speech but disregard the context. This paper investigates the role of context in the annotation and detection of online hate and counter speech, where context is defined as the preceding comment in a conversation thread. We created a context-aware dataset for a 3-way classification task on Reddit comments: hate speech, counter speech, or neutral. Our analyses indicate that context is critical to identify hate and counter speech: human judgments change for most comments depending on whether we show annotators the context. A linguistic analysis draws insights into the language people use to express hate and counter speech. Experimental results show that neural networks obtain significantly better results if context is taken into account. We also present qualitative error analyses shedding light into (a) when and why context is beneficial and (b) the remaining errors made by our best model when context is taken into account.

Machine Learning & Human-Centered NLP 2

09:15-10:15 (Columbia D)

[TACL] **Uncertainty Estimation and Reduction of Pre-trained Models for Text Regression**
Yuxia Wang, Daniel Beck, Timothy Baldwin and Karin Verspoor 09:15-09:30 (Columbia D)
State-of-the-art classification and regression models are often not well calibrated, and cannot reliably provide uncertainty estimates, limiting their utility in safety-critical applications such as clinical decision making. While recent work has focused on calibration of classifiers, there is almost no work in NLP on calibration in a regression setting. In this paper, we quantify the calibration of pre-trained language models for text regression, both intrinsically and extrinsically. We further apply uncertainty estimates to augment training data in low-resource domains. Our experiments on three regression tasks in both self-training and active-learning settings show that uncertainty estimation can be used to increase overall performance and enhance model generalisation.

[TACL] **Heterogeneous Supervised Topic Models**
Dhanya Sridhar, Hal Daumé III and David Meir Blei 09:30-09:45 (Columbia D)
Researchers in the social sciences are often interested in the relationship between text and an outcome of interest, where the goal is to both uncover latent patterns in the text and predict outcomes for unseen texts. To this end, this paper develops the heterogeneous supervised topic models (HSTM), a probabilistic approach to text analysis and prediction. HSTMs posit a joint model of text and outcomes to find heterogeneous patterns that help with both text analysis and prediction. The main benefit of HSTMs is that they capture heterogeneity in the relationship between text and the outcome across latent topics. To fit HSTMs, we develop a variational inference algorithm based on the auto-encoding variational Bayes framework. We study the performance of HSTMs on eight datasets and find that they consistently outperform related methods, including fine-tuned black box models. Finally, we apply HSTMs to analyze news articles labeled with pro- or anti-tone. We find evidence of differing language used to signal a pro- and anti-tone.

Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks
Paul Rottger, Bertie Vidgen, Dirk Hovy and Janet B. Pierrehumbert 09:45-10:00 (Columbia D)
Labelled data is the foundation of most natural language processing tasks. However, labelling data is difficult and there often are diverse valid beliefs about what the correct data labels should be. So far, dataset creators have acknowledged annotator subjectivity, but rarely actively managed it in the annotation process. This has led to partly-subjective datasets that fail to serve a clear downstream use. To address this issue, we propose two contrasting paradigms for data annotation. The descriptive paradigm encourages annotator subjectivity, whereas the prescriptive paradigm discourages it. Descriptive annotation allows for the surveying and modelling of different beliefs, whereas prescriptive annotation enables the training of models that consistently apply one belief. We discuss benefits and challenges in implementing both paradigms, and argue that dataset creators should explicitly aim for one or the other to facilitate the intended use of their dataset. Lastly, we conduct an annotation experiment using hate speech data that illustrates the contrast between the two paradigms.

On the Machine Learning of Ethical Judgments from Natural Language
Zeerak Talat, Hagen Blix, Josef Valvoda, Maya Indira Ganesh, Ryan Cotterell and Adina Williams 10:00-10:15 (Columbia D)
Ethics is one of the longest standing intellectual endeavors of humanity. In recent years, the fields of AI and NLP have attempted to address issues of harmful outcomes in machine learning systems that are made to interface with humans. One recent approach in this vein is the construction of NLP morality models that can take in arbitrary text and output a moral judgment about the situation described. In this work, we offer a critique of such NLP methods for automating ethical decision-making. Through an audit of recent work on computational approaches for predicting morality, we examine the broader issues that arise from such efforts. We conclude with a discussion of how machine ethics could usefully proceed in NLP, by focusing on current and near-future uses of technology, in a way that centers around transparency, democratic values, and allows for straightforward accountability.

Machine Translation 3

09:15-10:15 (Elwha A)

SURF: Semantic-level Unsupervised Reward Function for Machine Translation
Atijit Anuchitanukul and Julia Ive 09:15-09:30 (Elwha A)
The performance of Reinforcement Learning (RL) for natural language tasks including Machine Translation (MT) is crucially dependent on the reward formulation. This is due to the intrinsic difficulty of the task in the high-dimensional discrete action space as well as the sparseness of the standard reward functions defined for limited set of ground-truth sequences biased towards singular lexical choices. To address this issue, we formulate SURF, a maximally dense semantic-level unsupervised reward function which mimics human evaluation by considering both sentence fluency and semantic similarity. We demonstrate the strong potential of SURF to leverage a family of Actor-Critic Transformer-based Architectures with synchronous and asynchronous multi-agent variants. To tackle the problem of large action-state spaces, each agent is equipped with unique exploration strategies, promoting diversity during its exploration of the hypothesis space. When BLEU scores are compared, our dense unsupervised reward outperforms the standard sparse reward by 2% on average for in- and out-of-domain settings.

Reducing Disambiguation Biases in NMT by Leveraging Explicit Word Sense Information
Niccolò Campolungo, Tommaso Pasini, Denis Emelin and Roberto Navigli 09:30-09:45 (Elwha A)
Recent studies have shed some light on a common pitfall of Neural Machine Translation (NMT) models, stemming from their struggle to disambiguate polysemous words without lapsing into their most frequently occurring senses in the training corpus. In this paper, we first provide a novel approach for automatically creating high-precision sense-annotated parallel corpora, and then put forward a specifically tailored fine-tuning strategy for exploiting these sense annotations during training without introducing any additional requirement at inference time. The use of explicit senses proved to be beneficial to reduce the disambiguation bias of a baseline NMT model, while, at the same time, leading our system to attain higher BLEU scores than its vanilla counterpart in 3 language pairs.

Jam or Cream First? Modeling Ambiguity in Neural Machine Translation with SCONES
Felix Stahlberg and Shankar Kumar 09:45-10:00 (Elwha A)
The softmax layer in neural machine translation is designed to model the distribution over mutually exclusive tokens. Machine translation, however, is intrinsically uncertain: the same source sentence can have multiple semantically equivalent translations. Therefore, we propose to replace the softmax activation with a multi-label classification layer that can model ambiguity more effectively. We call our loss function Single-label Contrastive Objective for Non-Exclusive Sequences (SCONES). We show that the multi-label output layer can still be trained on single reference training data using the SCONES loss function. SCONES yields consistent BLEU score gains across six translation directions, particularly for medium-resource language pairs and small beam sizes. By using smaller beam sizes we can speed up inference by a factor of 3.9x and still match or improve the BLEU score obtained using softmax. Furthermore, we demonstrate that SCONES can be used to train NMT models that assign the highest probability to adequate translations, thus mitigating the "beam search curse”. Additional experiments on synthetic language pairs with varying levels of uncertainty suggest that the improvements from SCONES can be attributed to better handling of ambiguity.

Generating Authentic Adversarial Examples beyond Meaning-preserving with Doubly Round-trip Translation
Siyu Lai, Zhen Yang, Fandong Meng, Xue Zhang, Yufeng Chen, Jinan Xu and Jie Zhou 10:00-10:15 (Elwha A)
Generating adversarial examples for Neural Machine Translation (NMT) with single Round-Trip Translation (RTT) has achieved promising results by releasing the meaning-preserving restriction. However, a potential pitfall for this approach is that we cannot decide whether the generated examples are adversarial to the target NMT model or the auxiliary backward one, as the reconstruction error through the RTT can be related to either. To remedy this problem, we propose a new definition for NMT adversarial examples based on the Doubly Round-Trip Translation (DRTT). Specifically, apart from the source-target-source RTT, we also consider the target-source-target one, which is utilized to pick out the authentic adversarial examples for the target NMT model. Additionally, to enhance the robustness of the NMT model, we introduce the masked language models to construct bilingual adversarial pairs based on DRTT, which are used to train the NMT model directly. Extensive experiments on both the clean and noisy test sets (including the artificial and natural noise) show that our approach substantially improves the robustness of NMT models.

Dialogue and Interactive Systems 3

09:15-10:15 (Elwha B)

Design Challenges for a Multi-Perspective Search Engine
Sihao Chen, Siyi Liu, Xander Uyttendaele, Yi Zhang, William Bruno and Dan Roth 09:15-09:30 (Elwha B)
Many users turn to document retrieval systems (e.g. search engines) to seek answers to controversial or open-ended questions. However, classical document retrieval systems fall short at delivering users a set of direct and diverse responses in such cases, which requires identifying responses within web documents in the context of the query, and aggregating the responses based on their different perspectives.

The goal of this work is to survey and study the user information needs for building a multi-perspective search engine of such. We examine the challenges of synthesizing such language understanding objectives with document retrieval, and study a new *perspective-oriented* document retrieval paradigm. We discuss and assess the inherent natural language understanding challenges one needs to address in order to achieve the goal. Following the design challenges and principles, we propose and evaluate a practical prototype pipeline system. We use the prototype system to conduct a user survey in order to assess the utility of our paradigm, as well as understanding the user information needs when issuing controversial and open-ended queries to a search engine.

You Don’t Know My Favorite Color: Preventing Dialogue Representations from Revealing Speakers’ Private Personas
Haoran Li, Yangqiu Song and Lixin Fan 09:30-09:45 (Elwha B)
Social chatbots, also known as chit-chat chatbots, evolve rapidly with large pretrained language models. Despite the huge progress, privacy concerns have arisen recently: training data of large language models can be extracted via model inversion attacks. On the other hand, the datasets used for training chatbots contain many private conversations between two individuals. In this work, we further investigate the privacy leakage of the hidden states of chatbots trained by language modeling which has not been well studied yet. We show that speakers’ personas can be inferred through a simple neural network with high accuracy. To this end, we propose effective defense objectives to protect persona leakage from hidden states. We conduct extensive experiments to demonstrate that our proposed defense objectives can greatly reduce the attack accuracy from 37.6% to 0.5%. Meanwhile, the proposed objectives preserve language models’ powerful generation ability.

Unsupervised Slot Schema Induction for Task-oriented Dialog
Dian Yu, Mingqiu Wang, Yuan Cao, Izhak Shafran, Laurent El Shafey and Hagen Soltau 09:45-10:00 (Elwha B)
Carefully-designed schemas describing how to collect and annotate dialog corpora are a prerequisite towards building task-oriented dialog systems. In practical applications, manually designing schemas can be error-prone, laborious, iterative, and slow, especially when the schema is complicated. To alleviate this expensive and time consuming process, we propose an unsupervised approach for slot schema induction from unlabeled dialog corpora. Leveraging in-domain language models and unsupervised parsing structures, our data-driven approach extracts candidate slots without constraints, followed by coarse-to-fine clustering to induce slot types. We compare our method against several strong supervised baselines, and show significant performance improvement in slot schema induction on MultiWoz and SGD datasets. We also demonstrate the effectiveness of induced schemas on downstream applications including dialog state tracking and response generation.

CHAI: A CHatbot AI for Task-Oriented Dialogue with Offline Reinforcement Learning
Siddharth Verma, Justin Fu, Sherry Yang and Sergey Levine 10:00-10:15 (Elwha B)
Conventionally, generation of natural language for dialogue agents may be viewed as a statistical learning problem: determine the patterns in human-provided data and generate appropriate responses with similar statistical properties. However, dialogue can also be regarded as a goal directed process, where speakers attempt to accomplish a specific task. Reinforcement learning (RL) algorithms are designed specifically for solving such goal-directed problems, but the most direct way to apply RL, through trial-and-error learning in human conversations, is costly.

In this paper, we study how offline reinforcement learning can instead be used to train dialogue agents entirely using static datasets collected from human speakers. Our experiments show that recently developed offline RL methods can be combined with language models to yield realistic dialogue agents that better accomplish task goals.

Machine Learning for NLP 3

09:15-10:15 (Quinault)

Prompt Waywardness: The Curious Case of Discretized Interpretation of Continuous Prompts
Daniel Khashabi, Xinxi Lyu, Sewon Min, Lianhui Qin, Kyle Richardson, Sean Welleck, Hannaneh Hajishirzi, Tushar Khot, Ashish Sabharwal, Sameer Singh and Yejin Choi 09:15-09:30 (Quinault)
Fine-tuning continuous prompts for target tasks has recently emerged as a compact alternative to full model fine-tuning. Motivated by these promising results, we investigate the feasibility of extracting a discrete (textual) interpretation of continuous prompts that is faithful to the problem they solve. In practice, we observe a "wayward" behavior between the task solved by continuous prompts and their nearest neighbor discrete projections: We can find continuous prompts that solve a task while being projected to an arbitrary text (e.g., definition of a different or even a contradictory task), while being within a very small (2%) margin of the best continuous prompt of the same size for the task. We provide intuitions behind this odd and surprising behavior, as well as extensive empirical analyses quantifying the effect of various parameters. For instance, for larger model sizes we observe higher waywardness, i.e, we can find prompts that more closely map to any arbitrary text with a smaller drop in accuracy. These findings have important implications relating to the difficulty of faithfully interpreting continuous prompts and their generalization across models and tasks, providing guidance for future progress in prompting language models.

MetaICL: Learning to Learn In Context
Sewon Min, Mike Lewis, Luke Zettlemoyer and Hannaneh Hajishirzi 09:30-09:45 (Quinault)
We introduce MetaICL (Meta-training for In-Context Learning), a new meta-training framework for few-shot learning where a pretrained language model is tuned to do in-context learning on a large set of training tasks. This meta-training enables the model to more effectively learn a new task in context at test time, by simply conditioning on a few training examples with no parameter updates or task-specific templates. We experiment on a large, diverse collection of tasks consisting of 142 NLP datasets including classification, question answering, natural language inference, paraphrase detection and more, across seven different meta-training/target splits. MetaICL outperforms a range of baselines including in-context learning without meta-training and multi-task learning followed by zero-shot transfer. We find that the gains are particularly significant for target tasks that have domain shifts from the meta-training tasks, and that using a diverse set of the meta-training tasks is key to improvements. We also show that MetaICL approaches (and sometimes beats) the performance of models fully finetuned on the target task training data, and outperforms much bigger models with nearly 8x parameters.

Learning To Retrieve Prompts for In-Context Learning
Ohad Rubin, Jonathan Herzig and Jonathan Berant 09:45-10:00 (Quinault)
In-context learning is a recent paradigm in natural language understanding, where a large pre-trained language model (LM) observes a test instance and a few training examples as its input, and directly decodes the output without any update to its parameters. However, performance has been shown to strongly depend on the selected training examples (termed prompts). In this work, we propose an efficient method for retrieving prompts for in-context learning using annotated data and an LM. Given an input-output pair, we estimate the probability of the output given the input and a candidate training example as the prompt, and label training examples as positive or negative based on this probability. We then train an efficient dense retriever from this data, which is used to retrieve training examples as prompts at test time. We evaluate our approach on three sequence-to-sequence tasks where language utterances are mapped to meaning representations, and find that it substantially outperforms prior work and multiple baselines across the board.

IDPG: An Instance-Dependent Prompt Generation Method
Zhuofeng Wu, Sinong Wang, Jiatao Gu, Rui Hou, Yuxiao Dong, V.G.Vinod Vydiswaran and Hao Ma 10:00-10:15 (Quinault)
Prompt tuning is a new, efficient NLP transfer learning paradigm that adds a task-specific prompt in each input instance during the model training stage. It freezes the pre-trained language model and only optimizes a few task-specific prompts. In this paper, we propose a conditional prompt generation method to generate prompts for each input instance, referred to as the Instance-Dependent Prompt Generation (IDPG). Unlike traditional prompt tuning methods that use a fixed prompt, IDPG introduces a lightweight and trainable component to generate prompts based on each input sentence. Extensive experiments on ten natural language understanding (NLU) tasks show that the proposed strategy consistently outperforms various prompt tuning baselines and is on par with other efficient transfer learning methods such as Compacter while tuning far fewer model parameters.

Virtual Poster Q&A Session 4

09:15-10:15 (702 Clearwater)

Cooperative Self-training of Machine Reading Comprehension
Hongyin Luo, Shang-Wen Li, Mingye Gao, Seunghak Yu and James R. Glass 09:15-10:15 (702 Clearwater)
Pretrained language models have significantly improved the performance of downstream language understanding tasks, including extractive question answering, by providing high-quality contextualized word embeddings. However, training question answering models still requires large amounts of annotated data for specific domains. In this work, we propose a cooperative self-training framework, RGX, for automatically generating more non-trivial question-answer pairs to improve model performance. RGX is built upon a masked answer extraction task with an interactive learning environment containing an answer entity Recognizer, a question Generator, and an answer eXtractor. Given a passage with a masked entity, the generator generates a question around the entity, and the extractor is trained to extract the masked entity with the generated question and raw texts. The framework allows the training of question generation and answering models on any text corpora without annotation. We further leverage a self-training technique to improve the performance of both question generation and answer extraction models. Experiment results show that RGX outperforms the state-of-the-art (SOTA) pretrained language models and transfer learning approaches on standard question-answering benchmarks, and yields the new SOTA performance under given model size and transfer learning settings.

Ask Me Anything in Your Native Language
Nikita Sorokin, Dmitry Abulkhanov, Irina Piontkovskaya and Valentin Malykh 09:15-10:15 (702 Clearwater)

Cross-lingual question answering is a thriving field in the modern world, helping people to search information on the web more efficiently. One of the important scenarios is to give an answer even there is no answer in the language a person asks a question with. We present a novel approach based on single encoder for query and passage for retrieval from multi-lingual collection, together with cross-lingual generative reader. It achieves a new state of the art in both retrieval and end-to-end tasks on the XOR TyDi dataset outperforming the previous results up to 10% on several languages. We find that our approach can be generalized to more than 20 languages in zero-shot approach and outperform all previous models by 12%.

Yes, No or IDK: The Challenge of Unanswerable Yes/No Questions
Elior Sulem, Jamaal Hay and Dan Roth 09:15-10:15 (702 Clearwater)
The Yes/No QA task (Clark et al., 2019) consists of "Yes" or "No" questions about a given context. However, in realistic scenarios, the information provided in the context is not always sufficient in order to answer the question. For example, given the context "She married a lawyer from New-York.", we don't know whether the answer to the question "Did she marry in New York?" is "Yes" or "No". In this paper, we extend the Yes/No QA task, adding questions with an IDK answer, and show its considerable difficulty compared to the original 2-label task. For this purpose, we (i) enrich the BoolQ dataset (Clark et al., 2019) to include unanswerable questions and (ii) create out-of-domain test sets for the Yes/No/IDK QA task. We study the contribution of training on other Natural Language Understanding tasks. We focus in particular on Extractive QA (Rajpurkar et al., 2018) and Recognizing Textual Entailments (RTE; Dagan et al., 2013), analyzing the differences between 2 and 3 labels using the new data.

DREAM: Improving Situational QA by First Elaborating the Situation
Yuling Gu, Bhavana Dalvi and Peter Clark 09:15-10:15 (702 Clearwater)
When people answer questions about a specific situation, e.g., "I cheated on my mid-term exam last week. Was that wrong?", cognitive science suggests that they form a mental picture of that situation before answering. While we do not know how language models (LMs) answer such questions, we conjecture that they may answer more accurately if they are also provided with additional details about the question situation, elaborating the "scene". To test this conjecture, we train a new model, DREAM, to answer questions that elaborate the scenes that situated questions are about, and then provide those elaborations as additional context to a question-answering (QA) model. We find that DREAM is able to create better scene elaborations (more accurate, useful, and consistent) than a representative state-of-the-art, zero-shot model (Macaw). We also find that using the scene elaborations as additional context improves the answer accuracy of a downstream QA system, including beyond that obtainable by simply further fine-tuning the QA system on DREAM's training data. These results suggest that adding focused elaborations about a situation can improve a system's reasoning about it, and may serve as an effective way of injecting new scenario-based knowledge into QA models. Finally, our approach is dataset-neutral; we observe improved QA performance across different models, with even bigger gains on models with fewer parameters.

OPERA: Operation-Pivoted Discrete Reasoning over Text
Yongwei Zhou, Junwei Bao, Chaoqun Duan, Haipeng Sun, Jiahui Liang, Yifan Wang, Jing Zhao, Youzheng Wu, Xiaodong He and Tiejun Zhao
09:15-10:15 (702 Clearwater)
Machine reading comprehension (MRC) that requires discrete reasoning involving symbolic operations, e.g., addition, sorting, and counting, is a challenging task. According to this nature, semantic parsing-based methods predict interpretable but complex logical forms. However, logical form generation is nontrivial and even a little perturbation in a logical form will lead to wrong answers. To alleviate this issue, multi-predictor -based methods are proposed to directly predict different types of answers and achieve improvements. However, they ignore the utilization of symbolic operations and encounter a lack of reasoning ability and interpretability. To inherit the advantages of these two types of methods, we propose OPERA, an operation-pivoted discrete reasoning framework, where lightweight symbolic operations (compared with logical forms) as neural modules are utilized to facilitate the reasoning ability and interpretability. Specifically, operations are first selected and then softly executed to simulate the answer reasoning procedure. Extensive experiments on both DROP and RACENum datasets show the reasoning ability of OPERA. Moreover, further analysis verifies its interpretability.

TIE: Topological Information Enhanced Structural Reading Comprehension on Web Pages
Zihan Zhao, Lu Chen, Ruisheng Cao, Hongshen Xu, Xingyu Chen and Kai Yu 09:15-10:15 (702 Clearwater)
Recently, the structural reading comprehension (SRC) task on web pages has attracted increasing research interests. Although previous SRC work has leveraged extra information such as HTML tags or XPaths, the informative topology of web pages is not effectively exploited. In this work, we propose a Topological Information Enhanced model (TIE), which transforms the token-level task into a tag-level task by introducing a two-stage process (i.e. node locating and answer refining). Based on that, TIE integrates Graph Attention Network (GAT) and Pre-trained Language Model (PLM) to leverage the topological information of both logical structures and spatial structures. Experimental results demonstrate that our model outperforms strong baselines and achieves state-of-the-art performances on the web-based SRC benchmark WebSRC at the time of writing. The code of TIE will be publicly available at https://github.com/X-LANCE/TIE.

Long Context Question Answering via Supervised Contrastive Learning
Avi Caciularu, Ido Dagan, Jacob Goldberger and Arman Cohan 09:15-10:15 (702 Clearwater)
Long-context question answering (QA) tasks require reasoning over a long document or multiple documents. Addressing these tasks often benefits from identifying a set of evidence spans (e.g., sentences), which provide supporting evidence for answering the question. In this work, we propose a novel method for equipping long-context QA models with an additional sequence-level objective for better identification of the supporting evidence. We achieve this via an additional contrastive supervision signal in finetuning, where the model is encouraged to explicitly discriminate supporting evidence sentences from negative ones by maximizing question-evidence similarity. The proposed additional loss exhibits consistent improvements on three different strong long-context transformer models, across two challenging question answering benchmarks – HotpotQA and QAsper.

Dynamic Multistep Reasoning based on Video Scene Graph for Video Question Answering
Jianguo Mao, Wenbin Jiang, Xiangdong Wang, Zhifan Feng, Yajuan Lyu, Hong Liu and Yong Zhu 09:15-10:15 (702 Clearwater)
Existing video question answering (video QA) models lack the capacity for deep video understanding and flexible multistep reasoning. We propose for video QA a novel model which performs dynamic multistep reasoning between questions and videos. It creates video semantic representation based on the video scene graph composed of semantic elements of the video and semantic relations among these elements. Then, it performs multistep reasoning for better answer decision between the representations of the question and the video, and dynamically integrate the reasoning results. Experiments show the significant advantage of the proposed model against previous methods in accuracy and interpretability. Against the existing state-of-the-art model, the proposed model dramatically improves more than $4\%/3.1\%/2\%$ on the three widely used video QA datasets, MSRVTT-QA, MSRVTT multi-choice, and TGIF-QA, and displays better interpretability by backtracing along with the attention mechanisms to the video scene graphs.

A New Concept of Knowledge based Question Answering (KBQA) System for Multi-hop Reasoning
Yu Wang, v.srinivasan@samsung.com v.srinivasan@samsung.com and Hongxia Jin 09:15-10:15 (702 Clearwater)
Knowledge based question answering (KBQA) is a complex task for natural language understanding. Many KBQA approaches have been

proposed in recent years, and most of them are trained based on labeled reasoning path. This hinders the system's performance as many correct reasoning paths are not labeled as ground truth, and thus they cannot be learned. In this paper, we introduce a new concept of KBQA system which can leverage multiple reasoning paths' information and only requires labeled answer as supervision. We name it as **M**utliple **R**easoning **P**aths KB**QA** System (MRP-QA). We conduct experiments on several benchmark datasets containing both single-hop simple questions as well as muti-hop complex questions, including WebQuestionSP (WQSP), ComplexWebQuestion-1.1 (CWQ), and PathQuestion-Large (PQL), and demonstrate strong performance.

ProQA: Structural Prompt-based Pre-training for Unified Question Answering
Wanjun Zhong, Yifan Gao, Ning Ding, Yujia Qin, Zhiyuan Liu, Ming Zhou, Jiahai Wang, Jian Yin and Nan Duan 09:15-10:15 (702 Clearwater)
Question Answering (QA) is a longstanding challenge in natural language processing. Existing QA works mostly focus on specific question types, knowledge domains, or reasoning skills. The specialty in QA research hinders systems from modeling commonalities between tasks and generalization for wider applications. To address this issue, we present ProQA, a unified QA paradigm that solves various tasks through a single model. ProQA takes a unified structural prompt as the bridge and improves the QA-centric ability by structural prompt-based pre-training. Through a structurally designed prompt-based input schema, ProQA concurrently models the knowledge generalization for all QA tasks while keeping the knowledge customization for every specific QA task. Furthermore, ProQA is pre-trained with structural prompt-formatted large-scale synthesized corpus, which empowers the model with the commonly-required QA ability. Experimental results on 11 QA benchmarks demonstrate that ProQA consistently boosts performance on both full data fine-tuning, few-shot learning, and zero-shot testing scenarios. Furthermore, ProQA exhibits strong ability in both continual learning and transfer learning by taking the advantages of the structural prompt.

Multi-Hop Open-Domain Question Answering over Structured and Unstructured Knowledge
Yue Feng, Zhen Han, Mingming Sun and Ping Li 09:15-10:15 (702 Clearwater)
Open-domain question answering systems need to answer question of our interests with structured and unstructured information. However, existing approaches only select one source to generate answer or only conduct reasoning on structured information. In this paper, we propose a Document-Entity Heterogeneous Graph Network, referred to as DEHG, to effectively integrate different sources of information, and conduct reasoning on heterogeneous information. DEHG employs a graph constructor to integrate structured and unstructured information, a context encoder to represent nodes and question, a heterogeneous information reasoning layer to conduct multi-hop reasoning on both information sources, and an answer decoder to generate answers for the question. Experimental results on HybirdQA dataset show that DEHG outperforms the state-of-the-art methods.

Crake: Causal-Enhanced Table-Filler for Question Answering over Large Scale Knowledge Base
Minhao Zhang, Ruoyu Zhang, Yanzeng Li and Lei Zou 09:15-10:15 (702 Clearwater)
Semantic parsing solves knowledge base (KB) question answering (KBQA) by composing a KB query, which generally involves node extraction (NE) and graph composition (GC) to detect and connect related nodes in a query. Despite the strong causal effects between NE and GC, previous works fail to directly model such causalities in their pipeline, hindering the learning of subtask correlations. Also, the sequence-generation process for GC in previous works induces ambiguity and exposure bias, which further harms accuracy. In this work, we formalize semantic parsing into two stages. In the first stage (graph structure generation), we propose a causal-enhanced table-filler to overcome the issues in sequence-modelling and to learn the internal causalities. In the second stage (relation extraction), an efficient beam-search algorithm is presented to scale complex queries on large-scale KBs. Experiments on LC-QuAD 1.0 indicate that our method surpasses previous state-of-the-arts by a large margin (17%) while remaining time and space efficiency.

Incorporating Centering Theory into Neural Coreference Resolution
Haixia Chai and Michael Strube 09:15-10:15 (702 Clearwater)
In recent years, transformer-based coreference resolution systems have achieved remarkable improvements on the CoNLL dataset. However, how coreference resolvers can benefit from discourse coherence is still an open question. In this paper, we propose to incorporate centering transitions derived from centering theory in the form of a graph into a neural coreference model. Our method improves the performance over the SOTA baselines, especially on pronoun resolution in long documents, formal well-structured text, and clusters with scattered mentions.

ALLSH: Active Learning Guided by Local Sensitivity and Hardness
Shujian Zhang, Chengyue Gong, Xingchao Liu, Pengcheng He, Weizhu Chen and Mingyuan Zhou 09:15-10:15 (702 Clearwater)
Active learning, which effectively collects informative unlabeled data for annotation, reduces the demand for labeled data. In this work, we propose to retrieve unlabeled samples with a local sensitivity and hardness-aware acquisition function. The proposed method generates data copies through local perturbations and selects data points whose predictive likelihoods diverge the most from their copies. We further empower our acquisition function by injecting the select-worst case perturbation. Our method achieves consistent gains over the commonly used active learning strategies in various classification tasks. Furthermore, we observe consistent improvements over the baselines on the study of prompt selection in prompt-based few-shot learning. These experiments demonstrate that our acquisition guided by local sensitivity and hardness can be effective and beneficial for many NLP tasks.

Easy Adaptation to Mitigate Gender Bias in Multilingual Text Classification
Xiaolei Huang 09:15-10:15 (702 Clearwater)
Existing approaches to mitigate demographic biases evaluate on monolingual data, however, multilingual data has not been examined. In this work, we treat the gender as domains (e.g., male vs. female) and present a standard domain adaptation model to reduce the gender bias and improve performance of text classifiers under multilingual settings. We evaluate our approach on two text classification tasks, hate speech detection and rating prediction, and demonstrate the effectiveness of our approach with three fair-aware baselines.

Socially Aware Bias Measurements for Hindi Language Representations
Vijit Malik, Sunipa Dev, Akihiro Nishi, Nanyun Peng and Kai-Wei Chang 09:15-10:15 (702 Clearwater)
Language representations are an efficient tool used across NLP, but they are strife with encoded societal biases. These biases are studied extensively, but with a primary focus on English language representations and biases common in the context of Western society. In this work, we investigate the biases present in Hindi language representations such as caste and religion associated biases. We demonstrate how biases are unique to specific language representations based on the history and culture of the region they are widely spoken in, and also how the same societal bias (such as binary gender associated biases) when investigated across languages is encoded by different words and text spans. With this work, we emphasize on the necessity of social-awareness along with linguistic and grammatical artefacts when modeling language representations, in order to understand the biases encoded.

Recognition of They/Them as Singular Personal Pronouns in Coreference Resolution
Connor Baumler and Rachel Rudinger 09:15-10:15 (702 Clearwater)
As using they/them as personal pronouns becomes increasingly common in English, it is important that coreference resolution systems work

as well for individuals who use personal "they” as they do for those who use gendered personal pronouns. We introduce a new benchmark for coreference resolution systems which evaluates singular personal “they” recognition. Using these WinoNB schemas, we evaluate a number of publicly available coreference resolution systems and confirm their bias toward resolving “they” pronouns as plural.

Deconstructing NLG Evaluation: Evaluation Practices, Assumptions, and Their Implications
Kaitlyn Zhou, Su Lin Blodgett, Adam Trischler, Hal Daumé III, Kaheer Suleman and Alexandra Olteanu 09:15-10:15 (702 Clearwater)
There are many ways to express similar things in text, which makes evaluating natural language generation (NLG) systems difficult. Compounding this difficulty is the need to assess varying quality criteria depending on the deployment setting. While the landscape of NLG evaluation has been well-mapped, practitioners’ goals, assumptions, and constraints—which inform decisions about what, when, and how to evaluate—are often partially or implicitly stated, or not stated at all. Combining a formative semi-structured interview study of NLG practitioners (N=18) with a survey study of a broader sample of practitioners (N=61), we surface goals, community practices, assumptions, and constraints that shape NLG evaluations, examining their implications and how they embody ethical considerations.

BanglaBERT: Language Model Pretraining and Benchmarks for Low-Resource Language Understanding Evaluation in Bangla
Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman and Rifat Shahriyar 09:15-10:15 (702 Clearwater)
In this work, we introduce BanglaBERT, a BERT-based Natural Language Understanding (NLU) model pretrained in Bangla, a widely spoken yet low-resource language in the NLP literature. To pretrain BanglaBERT, we collect 27.5 GB of Bangla pretraining data (dubbed ’Bangla2B+’) by crawling 110 popular Bangla sites. We introduce two downstream task datasets on natural language inference and question answering and benchmark on four diverse NLU tasks covering text classification, sequence labeling, and span prediction. In the process, we bring them under the first-ever Bangla Language Understanding Benchmark (BLUB). BanglaBERT achieves state-of-the-art results outperforming multilingual and monolingual models. We are making the models, datasets, and a leaderboard publicly available at `https://github.com/csebuetnlp/banglabert` to advance Bangla NLP.

EVI: Multilingual Spoken Dialogue Tasks and Dataset for Knowledge-Based Enrolment, Verification, and Identification
Georgios P. Spithourakis, Ivan Vulić, Michał Lis, Inigo Casanueva and Paweł Budzianowski 09:15-10:15 (702 Clearwater)
Knowledge-based authentication is crucial for task-oriented spoken dialogue systems that offer personalised and privacy-focused services. Such systems should be able to enrol (E), verify (V), and identify (I) new and recurring users based on their personal information, e.g. postcode, name, and date of birth. In this work, we formalise the three authentication tasks and their evaluation protocols, and we present EVI, a challenging spoken multilingual dataset with 5,506 dialogues in English, Polish, and French. Our proposed models set the first competitive benchmarks, explore the challenges of multilingual natural language processing of spoken dialogue, and set directions for future research.

A Shoulder to Cry on: Towards A Motivational Virtual Assistant for Assuaging Mental Agony
Tulika Saha, Saichethan Miriyala Reddy, Anindya Sundar Das, Sriparna Saha and Pushpak Bhattacharyya 09:15-10:15 (702 Clearwater)
Mental Health Disorders continue plaguing humans worldwide. Aggravating this situation is the severe shortage of qualified and competent mental health professionals (MHPs), which underlines the need for developing Virtual Assistants (VAs) that can *assist* MHPs. The data+ML for automation can come from platforms that allow visiting and posting messages in peer-to-peer anonymous manner for sharing their experiences (frequently stigmatized) and seeking support. In this paper, we propose a VA that can act as the first point of contact and comfort for mental health patients. We curate a dataset, Motivational VA: MotiVAte comprising of 7k dyadic conversations collected from a peer-to-peer support platform. The system employs two mechanisms: (i) Mental Illness Classification: an attention based BERT classifier that outputs the mental disorder category out of the 4 categories, viz., Major Depressive Disorder (MDD), Anxiety, Obsessive Compulsive Disorder (OCD) and Post-traumatic Stress Disorder (PTSD), based on the input ongoing dialog between the support seeker and the VA; and (ii) Mental Illness Conditioned Motivational Dialogue Generation (MI-MDG): a sentiment driven Reinforcement Learning (RL) based motivational response generator. The empirical evaluation demonstrates the system capability by way of outperforming several baselines.

TWEETSPIN: Fine-grained Propaganda Detection in Social Media Using Multi-View Representations
Prashanth Vijayaraghavan and Soroush Vosoughi 09:15-10:15 (702 Clearwater)
Recently, several studies on propaganda detection have involved document and fragment-level analyses of news articles. However, there are significant data and modeling challenges dealing with fine-grained detection of propaganda on social media. In this work, we present TWEETSPIN, a dataset containing tweets that are weakly annotated with different fine-grained propaganda techniques, and propose a neural approach to detect and categorize propaganda tweets across those fine-grained categories. These categories include specific rhetorical and psychological techniques, ranging from leveraging emotions to using logical fallacies. Our model relies on multi-view representations of the input tweet data to (a) extract different aspects of the input text including the context, entities, their relationships, and external knowledge; (b) model their mutual interplay; and (c) effectively speed up the learning process by requiring fewer training examples. Our method allows for representation enrichment leading to better detection and categorization of propaganda on social media. We verify the effectiveness of our proposed method on TWEETSPIN and further probe how the implicit relations between the views impact the performance. Our experiments show that our model is able to outperform several benchmark methods and transfer the knowledge to relatively low-resource news domains.

Paragraph-based Transformer Pre-training for Multi-Sentence Inference
Luca Di Liello, Siddhant Garg, Luca Soldaini and Alessandro Moschitti 09:15-10:15 (702 Clearwater)
Inference tasks such as answer sentence selection (AS2) or fact verification are typically solved by fine-tuning transformer-based models as individual sentence-pair classifiers. Recent studies show that these tasks benefit from modeling dependencies across multiple candidate sentences jointly. In this paper, we first show that popular pre-trained transformers perform poorly when used for fine-tuning on multi-candidate inference tasks. We then propose a new pre-training objective that models the paragraph-level semantics across multiple input sentences. Our evaluation on three AS2 and one fact verification datasets demonstrates the superiority of our pre-training technique over the traditional ones for transformers used as joint models for multi-candidate inference tasks, as well as when used as cross-encoders for sentence-pair formulations of these tasks.

Few-Shot Semantic Parsing with Language Models Trained on Code
Richard Shin and Benjamin Van Durme 09:15-10:15 (702 Clearwater)
Large language models can perform semantic parsing with little training data, when prompted with in-context examples. It has been shown that this can be improved by formulating the problem as paraphrasing into canonical utterances, which casts the underlying meaning representation into a controlled natural language-like representation. Intuitively, such models can more easily output canonical utterances as they are closer to the natural language used for pre-training. Recently, models also pre-trained on code, like OpenAI Codex, have risen in prominence. For semantic parsing tasks where we map natural language into code, such models may prove more adept at it. In this paper, we test this hypothesis and find that Codex performs better on such tasks than equivalent GPT-3 models. We evaluate on Overnight and SMCalFlow and find that unlike GPT-3, Codex performs similarly when targeting meaning representations directly, perhaps because meaning representations are structured similar to code in these datasets.

EmRel: Joint Representation of Entities and Embedded Relations for Multi-triple Extraction
Benfeng Xu, Quan Wang, Yajuan Lyu, Yabing Shi, Yong Zhu, Jie Gao and Zhendong Mao 09:15-10:15 (702 Clearwater)
Multi-triple extraction is a challenging task due to the existence of informative inter-triple correlations, and consequently rich interactions across the constituent entities and relations. While existing works only explore entity representations, we propose to explicitly introduce *relation* representation, jointly represent it with entities, and novelly align them to identify valid triples. We perform comprehensive experiments on document-level relation extraction and joint entity and relation extraction along with ablations to demonstrate the advantage of the proposed method.

CompactIE: Compact Facts in Open Information Extraction
Farima Fatahi Bayat, Nikita Bhutani and H. Jagadish 09:15-10:15 (702 Clearwater)
A major drawback of modern neural OpenIE systems and benchmarks is that they prioritize high coverage of information in extractions over compactness of their constituents. This severely limits the usefulness of OpenIE extractions in many downstream tasks. The utility of extractions can be improved if extractions are compact and share constituents. To this end, we study the problem of identifying compact extractions with neural-based methods. We propose CompactIE, an OpenIE system that uses a novel pipelined approach to produce compact extractions with overlapping constituents. It first detects constituents of the extractions and then links them to build extractions. We train our system on compact extractions obtained by processing existing benchmarks. Our experiments on CaRB and Wire57 datasets indicate that CompactIE finds 1.5x-2x more compact extractions than previous systems, with high precision, establishing a new state-of-the-art performance in OpenIE.

Document-Level Relation Extraction with Sentences Importance Estimation and Focusing
Wang Xu, Kehai Chen, Lili Mou and Tiejun Zhao 09:15-10:15 (702 Clearwater)
Document-level relation extraction (DocRE) aims to determine the relation between two entities from a document of multiple sentences. Recent studies typically represent the entire document by sequence- or graph-based models to predict the relations of all entity pairs. However, we find that such a model is not robust and exhibits bizarre behaviors: it predicts correctly when an entire test document is fed as input, but errs when non-evidence sentences are removed. To this end, we propose a Sentence Importance Estimation and Focusing (SIEF) framework for DocRE, where we design a sentence importance score and a sentence focusing loss, encouraging DocRE models to focus on evidence sentences. Experimental results on two domains show that our SIEF not only improves overall performance, but also makes DocRE models more robust. Moreover, SIEF is a general framework, shown to be effective when combined with a variety of base DocRE models.

ITA: Image-Text Alignments for Multi-Modal Named Entity Recognition
Xinyu Wang, Min Gui, Yong Jiang, Zixia Jia, Nguyen Bach, Tao Wang, Zhongqiang Huang and Kewei Tu 09:15-10:15 (702 Clearwater)
Recently, Multi-modal Named Entity Recognition (MNER) has attracted a lot of attention. Most of the work utilizes image information through region-level visual representations obtained from a pretrained object detector and relies on an attention mechanism to model the interactions between image and text representations. However, it is difficult to model such interactions as image and text representations are trained separately on the data of their respective modality and are not aligned in the same space. As text representations take the most important role in MNER, in this paper, we propose **I**mage-**t**ext **A**lignments (ITA) to align image features into the textual space, so that the attention mechanism in transformer-based pretrained textual embeddings can be better utilized. ITA first aligns the image into regional object tags, image-level captions and optical characters as visual contexts, concatenates them with the input texts as a new cross-modal input, and then feeds it into a pretrained textual embedding model. This makes it easier for the attention module of a pretrained textual embedding model to model the interaction between the two modalities since they are both represented in the textual space. ITA further aligns the output distributions predicted from the cross-modal input and textual input views so that the MNER model can be more practical in dealing with text-only inputs and robust to noises from images. In our experiments, we show that ITA models can achieve state-of-the-art accuracy on multi-modal Named Entity Recognition datasets, even without image information.

Hierarchical Relation-Guided Type-Sentence Alignment for Long-Tail Relation Extraction with Distant Supervision
Yang Li, Guodong Long, Tao Shen and Jing Jiang 09:15-10:15 (702 Clearwater)
Distant supervision uses triple facts in knowledge graphs to label a corpus for relation extraction, leading to wrong labeling and long-tail problems. Some works use the hierarchy of relations for knowledge transfer to long-tail relations. However, a coarse-grained relation often implies only an attribute (e.g., domain or topic) of the distant fact, making it hard to discriminate relations based solely on sentence semantics. One solution is resorting to entity types, but open questions remain about how to fully leverage the information of entity types and how to align multi-granular entity types with sentences. In this work, we propose a novel model to enrich distantly-supervised sentences with entity types. It consists of (1) a pairwise type-enriched sentence encoding module injecting both context-free and -related backgrounds to alleviate sentence-level wrong labeling, and (2) a hierarchical type-sentence alignment module enriching a sentence with the triple fact's basic attributes to support long-tail relations. Our model achieves new state-of-the-art results in overall and long-tail performance on benchmarks.

Good Visual Guidance Make A Better Extractor: Hierarchical Visual Prefix for Multimodal Entity and Relation Extraction
Xiang Chen, Ningyu Zhang, Lei Li, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si and Huajun Chen 09:15-10:15 (702 Clearwater)
Multimodal named entity recognition and relation extraction (MNER and MRE) is a fundamental and crucial branch in information extraction. However, existing approaches for MNER and MRE usually suffer from error sensitivity when irrelevant object images incorporated in texts. To deal with these issues, we propose a novel Hierarchical Visual Prefix fusion NeTwork (HVPNeT) for visual-enhanced entity and relation extraction, aiming to achieve more effective and robust performance. Specifically, we regard visual representation as pluggable visual prefix to guide the textual representation for error insensitive forecasting decision. We further propose a dynamic gated aggregation strategy to achieve hierarchical multi-scaled visual features as visual prefix for fusion. Extensive experiments on three benchmark datasets demonstrate the effectiveness of our method, and achieve state-of-the-art performance.

Label Refinement via Contrastive Learning for Distantly-Supervised Named Entity Recognition
Huaiyuan Ying, Shengxuan Luo, Tiantian Dang and Sheng Yu 09:15-10:15 (702 Clearwater)
Distantly-supervised named entity recognition (NER) locates and classifies entities using only knowledge bases and unlabeled corpus to mitigate the reliance on human-annotated labels. The distantly annotated data suffer from the noise in labels, and previous works on DSNER have proved the importance of pre-refining distant labels with hand-crafted rules and extra existing semantic information. In this work, we explore the way to directly learn the distant label refinement knowledge by imitating annotations of different qualities and comparing these annotations in contrastive learning frameworks. the proposed distant label refinement model can give modified suggestions on distant data without additional supervised labels, and thus reduces the requirement on the quality of the knowledge bases. We perform extensive experiments and observe that recent and state-of-the-art DSNER methods gain evident benefits with our method.

Machine-in-the-Loop Rewriting for Creative Image Captioning
Vishakh Padmakumar and He He 09:15-10:15 (702 Clearwater)
Machine-in-the-loop writing aims to build models that assist humans to accomplish their writing tasks more effectively. Prior work has found that providing users a machine-written draft or sentence-level continuations has limited success since the generated text tends to deviate from

users' intention. To allow the user to retain control over the content, we train a rewriting model that, when prompted, modifies specified spans of text within the user's original draft to introduce descriptive and figurative elements in the text. We evaluate the model on its ability to collaborate with humans on the task of creative image captioning. On a user study through Amazon Mechanical Turk, our model is rated to be more helpful by users than a baseline infilling language model. In addition, third-party evaluation shows that users write more descriptive and figurative captions when collaborating with our model compared to completing the task alone. However, the improvement is not uniform across user groups: the model is more helpful to skilled users, which risks widening the gap between skilled and novice users, highlighting a need for careful, user-centric evaluation of interactive systems.

Twitter-COMMs: Detecting Climate, COVID, and Military Multimodal Misinformation
Giscard Biamby, Grace Luo, Trevor Darrell and Anna Rohrbach 09:15-10:15 (702 Clearwater)
Detecting out-of-context media, such as "miscaptioned" images on Twitter, is a relevant problem, especially in domains of high public significance. In this work we aim to develop defenses against such misinformation for the topics of Climate Change, COVID-19, and Military Vehicles. We first present a large-scale multimodal dataset with over 884k tweets relevant to these topics. Next, we propose a detection method, based on the state-of-the-art CLIP model, that leverages automatically generated hard image-text mismatches. While this approach works well on our automatically constructed out-of-context tweets, we aim to validate its usefulness on data representative of the real world. Thus, we test it on a set of human-generated fakes, created by mimicking in-the-wild misinformation. We achieve an 11% detection improvement in a high precision regime over a strong baseline. Finally, we share insights about our best model design and analyze the challenges of this emerging threat.

MCSE: Multimodal Contrastive Learning of Sentence Embeddings
Miaoran Zhang, Marius Mosbach, David Ifeoluwa Adelani, Michael A. Hedderich and Dietrich Klakow 09:15-10:15 (702 Clearwater)
Learning semantically meaningful sentence embeddings is an open problem in natural language processing. In this work, we propose a sentence embedding learning approach that exploits both visual and textual information via a multimodal contrastive objective. Through experiments on a variety of semantic textual similarity tasks, we demonstrate that our approach consistently improves the performance across various datasets and pre-trained encoders. In particular, combining a small amount of multimodal data with a large text-only corpus, we improve the state-of-the-art average Spearman's correlation by 1.7%. By analyzing the properties of the textual embedding space, we show that our model excels in aligning semantically similar sentences, providing an explanation for its improved performance.

Fine-grained Image Captioning with CLIP Reward
Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui and Mohit Bansal 09:15-10:15 (702 Clearwater)
Modern image captioning models are usually trained with text similarity objectives. However, since reference captions in public datasets often describe the most salient common objects, models trained with the text similarity objectives tend to ignore specific and detailed aspects of an image that distinguish it from others. Towards more descriptive and distinctive caption generation, we propose to use CLIP, a multimodal encoder trained on huge image-text pairs from the web, to calculate multi-modal similarity and use it as a reward function. We also propose a simple finetuning strategy of CLIP text encoder to improve grammar that does not require extra text annotation. This completely eliminates the need for reference captions during the reward computation. To comprehensively evaluate descriptive captions, we introduce FineCapEval, a new dataset for caption evaluation with fine-grained criteria: overall, background, object, relations. In our experiments on text-to-image retrieval and FineCapEval, the proposed CLIP-guided model generates more distinctive captions than the CIDEroptimized model. We also show that our unsupervised grammar finetuning of the CLIP text encoder alleviates the degeneration problem of the naive CLIP reward. Lastly, we show human analysis where the annotators strongly prefer CLIP reward to CIDEr and MLE objectives on diverse criteria.

CLEAR: Improving Vision-Language Navigation with Cross-Lingual, Environment-Agnostic Representations
Jialu Li, Hao Tan and Mohit Bansal 09:15-10:15 (702 Clearwater)
Vision-and-Language Navigation (VLN) tasks require an agent to navigate through the environment based on language instructions. In this paper, we aim to solve two key challenges in this task: utilizing multilingual instructions for improved instruction-path grounding and navigating through new environments that are unseen during training. To address these challenges, first, our agent learns a shared and visually-aligned cross-lingual language representation for the three languages (English, Hindi and Telugu) in the Room-Across-Room dataset. Our language representation learning is guided by text pairs that are aligned by visual information. Second, our agent learns an environment-agnostic visual representation by maximizing the similarity between semantically-aligned image pairs (with constraints on object-matching) from different environments. Our environment agnostic visual representation can mitigate the environment bias induced by low-level visual information. Empirically, on the Room-Across-Room dataset, we show that our multi-lingual agent gets large improvements in all metrics over the strong baseline model when generalizing to unseen environments with the cross-lingual language representation and the environment-agnostic visual representation. Furthermore, we show that our learned language and visual representations can be successfully transferred to the Room-to-Room and Cooperative Vision-and-Dialogue Navigation task, and present detailed qualitative and quantitative generalization and grounding analysis.

What kinds of errors do reference resolution models make and what can we learn from them?
Jorge Sánchez, Mauricio Mazuecos, Hernán Maina and Luciana Benotti 09:15-10:15 (702 Clearwater)
Referring resolution is the task of identifying the referent of a natural language expression, for example "the woman behind the other woman getting a massage". In this paper we investigate which are the kinds of referring expressions on which current transformer based models fail. Motivated by this analysis we identify the weakening of the spatial natural constraints as one of its causes and propose a model that aims to restore it. We evaluate our proposed model on different datasets for the task showing improved performance on the most challenging kinds of referring expressions. Finally we present a thorough analysis of the kinds errors that are improved by the new model and those that are not and remain future challenges for the task.

Negative Sample is Negative in Its Own Way: Tailoring Negative Sentences for Image-Text Retrieval
Zhihao Fan, Zhongyu Wei, Zejun Li, Siyuan Wang, Xuanjing Huang and Jianqing Fan 09:15-10:15 (702 Clearwater)
Matching model is essential for Image-Text Retrieval framework. Existing research usually train the model with a triplet loss and explore various strategy to retrieve hard negative sentences in the dataset. We argue that current retrieval-based negative sample construction approach is limited in the scale of the dataset thus fail to identify negative sample of high difficulty for every image. We propose our TAiloring neGative Sentences with Discrimination and Correction (TAGS-DC) to generate synthetic sentences automatically as negative samples. TAGS-DC is composed of masking and refilling to generate synthetic negative sentences with higher difficulty. To keep the difficulty during training, we mutually improve the retrieval and generation through parameter sharing. To further utilize fine-grained semantic of mismatch in the negative sentence, we propose two auxiliary tasks, namely word discrimination and word correction to improve the training. In experiments, we verify the effectiveness of our model on MS-COCO and Flickr30K compared with current state-of-the-art models and demonstrates its robustness and faithfulness in the further analysis.

RoViST: Learning Robust Metrics for Visual Storytelling
Eileen Wang, Caren Han and Josiah Poon 09:15-10:15 (702 Clearwater)

Visual storytelling (VST) is the task of generating a story paragraph that describes a given image sequence. Most existing storytelling approaches have evaluated their models using traditional natural language generation metrics like BLEU or CIDEr. However, such metrics based on n-gram matching tend to have poor correlation with human evaluation scores and do not explicitly consider other criteria necessary for storytelling such as sentence structure or topic coherence. Moreover, a single score is not enough to assess a story as it does not inform us about what specific errors were made by the model. In this paper, we propose 3 evaluation metrics sets that analyses which aspects we would look for in a good story: 1) visual grounding, 2) coherence, and 3) non-redundancy. We measure the reliability of our metric sets by analysing its correlation with human judgement scores on a sample of machine stories obtained from 4 state-of-the-arts models trained on the Visual Storytelling Dataset (VIST). Our metric sets outperforms other metrics on human correlation, and could be served as a learning based evaluation metric set that is complementary to existing rule-based metrics.

Detecting Narrative Elements in Informational Text
Effi Levi, Guy Mor, Tamir Sheafer and Shaul Rafael Shenhav 09:15-10:15 (702 Clearwater)
Automatic extraction of narrative elements from text, combining narrative theories with computational models, has been receiving increasing attention over the last few years. Previous works have utilized the oral narrative theory by Labov and Waletzky to identify various narrative elements in personal stories texts. Instead, we direct our focus to informational texts, specifically news stories.

We introduce NEAT (Narrative Elements AnnoTation) – a novel NLP task for detecting narrative elements in raw text. For this purpose, we designed a new multi-label narrative annotation scheme, better suited for informational text (e.g. news media), by adapting elements from the narrative theory of Labov and Waletzky (Complication and Resolution) and adding a new narrative element of our own (Success). We then used this scheme to annotate a new dataset of 2,209 sentences, compiled from 46 news articles from various category domains. We trained a number of supervised models in several different setups over the annotated dataset to identify the different narrative elements, achieving an average F_1 score of up to 0.77. The results demonstrate the holistic nature of our annotation scheme as well as its robustness to domain category.

Pretrained Models for Multilingual Federated Learning
Orion Weller, Marc Marone, Vladimir Braverman, Dawn Lawrie and Benjamin Van Durme 09:15-10:15 (702 Clearwater)
Since the advent of Federated Learning (FL), research has applied these methods to natural language processing (NLP) tasks. Despite a plethora of papers in FL for NLP, no previous works have studied how multilingual text impacts FL algorithms. Furthermore, multilingual text provides an interesting avenue to examine the impact of non-IID text (e.g. different languages) on FL in naturally occurring data. We explore three multilingual language tasks, language modeling, machine translation, and text classification using differing federated and non-federated learning algorithms. Our results show that using pretrained models reduces the negative effects of FL, helping them to perform near or better than centralized (no privacy) learning, even when using non-IID partitioning.

BAD-X: Bilingual Adapters Improve Zero-Shot Cross-Lingual Transfer
Marinela Parović, Goran Glavaš, Ivan Vulić and Anna Korhonen 09:15-10:15 (702 Clearwater)
Adapter modules enable modular and efficient zero-shot cross-lingual transfer, where current state-of-the-art adapter-based approaches learn specialized language adapters (LAs) for individual languages. In this work, we show that it is more effective to learn bilingual language pair adapters (BAs) when the goal is to optimize performance for a particular source-target transfer direction. Our novel BAD-X adapter framework trades off some modularity of dedicated LAs for improved transfer performance: we demonstrate consistent gains in three standard downstream tasks, and for the majority of evaluated low-resource languages.

Towards Debiasing Translation Artifacts
Koel Dutta Chowdhury, Rricha Jalota, Cristina España-Bonet and Josef Van Genabith 09:15-10:15 (702 Clearwater)
Cross-lingual natural language processing relies on translation, either by humans or machines, at different levels, from translating training data to translating test sets. However, compared to original texts in the same language, translations possess distinct qualities referred to as translationese. Previous research has shown that these translation artifacts influence the performance of a variety of cross-lingual tasks. In this work, we propose a novel approach to reducing translationese by extending an established bias-removal technique. We use the Iterative Null-space Projection (INLP) algorithm, and show by measuring classification accuracy before and after debiasing, that translationese is reduced at both sentence and word level. We evaluate the utility of debiasing translationese on a natural language inference (NLI) task, and show that by reducing this bias, NLI accuracy improves. To the best of our knowledge, this is the first study to debias translationese as represented in latent embedding space.

Opportunities for Human-centered Evaluation of Machine Translation Systems
Daniel J. Liebling, Katherine A Heller, Samantha Robertson and Wesley Deng 09:15-10:15 (702 Clearwater)
Machine translation models are embedded in larger user-facing systems. Although model evaluation has matured, evaluation at the systems level is still lacking. We review literature from both the translation studies and HCI communities about who uses machine translation and for what purposes. We emphasize an important difference in evaluating machine translation models versus the physical and cultural systems in which they are embedded. We then propose opportunities for improved measurement of user-facing translation systems. We pay particular attention to the need for design and evaluation to aid engendering trust and enhancing user agency in future machine translation systems.

Uncertainty-Aware Cross-Lingual Transfer with Pseudo Partial Labels
Shuo Lei, Xuchao Zhang, Jianfeng He, Fanglan Chen and Chang-Tien Lu 09:15-10:15 (702 Clearwater)
Large-scale multilingual pre-trained language models have achieved remarkable performance in zero-shot cross-lingual tasks. A recent study has demonstrated the effectiveness of self-learning-based approach on cross-lingual transfer, where only unlabeled data of target languages are required, without any efforts to annotate gold labels for target languages. However, it suffers from noisy training due to the incorrectly pseudo-labeled samples. In this work, we propose an uncertainty-aware Cross-Lingual Transfer framework with Pseudo-Partial-Label (CLTP)1 to maximize the utilization of unlabeled data by reducing the noise introduced in the training phase. To estimate pseudo-partial-label for each unlabeled data, we propose a novel estimation method, considering both prediction confidence and the limitation to the number of similar labels. Extensive experiments are conducted on two cross-lingual tasks, including Named Entity Recognition (NER) and Natural Language Inference (NLI) across 40 languages, which shows our method can outperform the baselines on both high-resource and low-resource languages, such as 6.9 on Kazakh (kk) and 5.2 Marathi (mr) for NER.

Multi-stage Distillation Framework for Cross-Lingual Semantic Similarity Matching
Kunbo Ding, Weijie Liu, Yuejian Fang, Zhe Zhao, Qi Ju, Xuefeng Yang, Rong Tian, Zhu Tao, Haoyan Liu, Han Guo, Xingyu Bai, Weiquan Mao, Yudong Li, Weigang Guo, Taiqiang Wu and Ningyuan Sun 09:15-10:15 (702 Clearwater)
Previous studies have proved that cross-lingual knowledge distillation can significantly improve the performance of pre-trained models for cross-lingual similarity matching tasks. However, the student model needs to be large in this operation. Otherwise, its performance will drop sharply, thus making it impractical to be deployed to memory-limited devices. To address this issue, we delve into cross-lingual knowledge distillation and propose a multi-stage distillation framework for constructing a small-size but high-performance cross-lingual model. In our framework, contrastive learning, bottleneck, and parameter recurrent strategies are delicately combined to prevent performance from being

compromised during the compression process. The experimental results demonstrate that our method can compress the size of XLM-R and MiniLM by more than 50%, while the performance is only reduced by about 1%.

Reference-free Summarization Evaluation via Semantic Correlation and Compression Ratio
Yizhu Liu, Qi Jia and Kenny Q. Zhu 09:15-10:15 (702 Clearwater)
A document can be summarized in a number of ways. Reference-based evaluation of summarization has been criticized for its inflexibility. The more sufficient the number of abstracts, the more accurate the evaluation results. However, it is difficult to collect sufficient reference summaries. In this paper, we propose a new automatic reference-free evaluation metric that compares semantic distribution between source document and summary by pretrained language models and considers summary compression ratio. The experiments show that this metric is more consistent with human evaluation in terms of coherence, consistency, relevance and fluency.

Data Augmentation for Low-Resource Dialogue Summarization
Yongtai Liu, Joshua Maynez, Gonçalo Simões and Shashi Narayan 09:15-10:15 (702 Clearwater)
We present DADS, a novel Data Augmentation technique for low-resource Dialogue Summarization. Our method generates synthetic examples by replacing sections of text from both the input dialogue and summary while preserving the augmented summary to correspond to a viable summary for the augmented dialogue. We utilize pretrained language models that produce highly likely dialogue alternatives while still being free to generate diverse alternatives. We applied our data augmentation method to the SAMSum dataset in low resource scenarios, mimicking real world problems such as chat, thread, and meeting summarization where large scale supervised datasets with human-written summaries are scarce. Through both automatic and human evaluations, we show that DADS shows strong improvements for low resource scenarios while generating topically diverse summaries without introducing additional hallucinations to the summaries.

OTExtSum: Extractive Text Summarisation with Optimal Transport
Peggy Tang, Kun Hu, Rui Yan, Lei Zhang, Junbin Gao and Zhiyong Wang 09:15-10:15 (702 Clearwater)
Extractive text summarisation aims to select salient sentences from a document to form a short yet informative summary. While learning-based methods have achieved promising results, they have several limitations, such as dependence on expensive training and lack of interpretability. Therefore, in this paper, we propose a novel non-learning-based method by for the first time formulating text summarisation as an Optimal Transport (OT) problem, namely Optimal Transport Extractive Summariser (OTExtSum). Optimal sentence extraction is conceptualised as obtaining an optimal summary that minimises the transportation cost to a given document regarding their semantic distributions. Such a cost is defined by the Wasserstein distance and used to measure the summary's semantic coverage of the original document. Comprehensive experiments on four challenging and widely used datasets - MultiNews, PubMed, BillSum, and CNN/DM demonstrate that our proposed method outperforms the state-of-the-art non-learning-based methods and several recent learning-based methods in terms of the ROUGE metric.

Exploring Neural Models for Query-Focused Summarization
Jesse Vig, Alexander Fabbri, Wojciech Maciej Kryscinski, Chien-Sheng Wu and Wenhao Liu 09:15-10:15 (702 Clearwater)
Query-focused summarization (QFS) aims to produce summaries that answer particular questions of interest, enabling greater user control and personalization. While recently released datasets, such as QMSum or AQuaMuSe, facilitate research efforts in QFS, the field lacks a comprehensive study of the broad space of applicable modeling methods. In this paper we conduct a systematic exploration of neural approaches to QFS, considering two general classes of methods: two-stage extractive-abstractive solutions and end-to-end models. Within those categories, we investigate existing models and explore strategies for transfer learning. We also present two modeling extensions that achieve state-of-the-art performance on the QMSum dataset, up to a margin of 3.38 ROUGE-1, 3.72 ROUGE2, and 3.28 ROUGE-L when combined with transfer learning strategies. Results from human evaluation suggest that the best models produce more comprehensive and factually consistent summaries compared to a baseline model. Code and checkpoints are made publicly available: https://github.com/salesforce/query-focused-sum.

Post-Training Dialogue Summarization using Pseudo-Paraphrasing
Qi Jia, Yizhu Liu, Haifeng Tang and Kenny Q. Zhu 09:15-10:15 (702 Clearwater)
Previous dialogue summarization techniques adapt large language models pretrained on the narrative text by injecting dialogue-specific features into the models. These features either require additional knowledge to recognize or make the resulting models harder to tune. To bridge the format gap between dialogues and narrative summaries in dialogue summarization tasks, we propose to post-train pretrained language models (PLMs) to rephrase from dialogue to narratives. After that, the model is fine-tuned for dialogue summarization as usual. Comprehensive experiments show that our approach significantly improves vanilla PLMs on dialogue summarization and outperforms other SOTA models by the summary quality and implementation costs.

TANet: Thread-Aware Pretraining for Abstractive Conversational Summarization
Ze Yang, Christian Wang, Zhoujin Tian, Wei Wu and Zhoujun Li 09:15-10:15 (702 Clearwater)
Although pre-trained language models (PLMs) have achieved great success and become a milestone in NLP, abstractive conversational summarization remains a challenging but less studied task. The difficulty lies in two aspects. One is the lack of large-scale conversational summary data. Another is that applying the existing pre-trained models to this task is tricky because of the structural dependence within the conversation and its informal expression, etc. In this work, we first build a large-scale (11M) pretraining dataset called RCSum, based on the multi-person discussions in the Reddit community. We then present TANet, a thread-aware Transformer-based network. Unlike the existing pre-trained models that treat a conversation as a sequence of sentences, we argue that the inherent contextual dependency among the utterances plays an essential role in understanding the entire conversation and thus propose two new techniques to incorporate the structural information into our model. The first is thread-aware attention which is computed by taking into account the contextual dependency within utterances. Second, we apply thread prediction loss to predict the relations between utterances. We evaluate our model on four datasets of real conversations, covering types of meeting transcripts, customer-service records, and forum threads. Experimental results demonstrate that TANet achieves a new state-of-the-art in terms of both automatic evaluation and human judgment.

Jointly Learning Guidance Induction and Faithful Summary Generation via Conditional Variational Autoencoders
Wang Xu and Tiejun Zhao 09:15-10:15 (702 Clearwater)
Abstractive summarization can generate high quality results with the development of the neural network. However, generating factual consistency summaries is a challenging task for abstractive summarization. Recent studies extract the additional information with off-the-shelf tools from the source document as a clue to guide the summary generation, which shows effectiveness to improve the faithfulness. Unlike these work, we present a novel framework based on conditional variational autoencoders, which induces the guidance information and generates the summary equipment with the guidance synchronously. Experiments on XSUM and CNNDM dataset show that our approach can generate relevant and fluent summaries which is more faithful than the existing state-of-the-art approaches, according to multiple factual consistency metrics.

Understanding Long Document with Different Position-Aware Attentions
Hai Pham, Guoxin Wang, Yijuan Lu, Dinei Florencio and Cha Zhang 09:15-10:15 (702 Clearwater)

Despite several successes in document understanding, the practical task for long document understanding is largely under-explored due to several challenges in computation and how to efficiently absorb long multimodal input. Most current transformer-based approaches only deal with short documents and employ solely textual information for attention due to its prohibitive computation and memory limit. To address those issues in long document understanding, we explore different approaches in handling 1D and new 2D position-aware attention with essentially shortened context. Experimental results show that our proposed models have the advantages for this task based on various evaluation metrics. Furthermore, our model makes changes only to the attention and thus can be easily used for any transformer-based architecture.

Dr. Livingstone, I presume? Polishing of foreign character identification in literary texts
Aleksandra Konovalova, Antonio Toral and Kristiina Taivalkoski-Shilov 09:15-10:15 (702 Clearwater)
Character identification is a key element for many narrative-related tasks. To implement it, the baseform of the name of the character (or lemma) needs to be identified, so different appearances of the same character in the narrative could be aligned. In this paper we tackle this problem in translated texts (English–Finnish translation direction), where the challenge regarding lemmatizing foreign names in an agglutinative language appears. To solve this problem, we present and compare several methods. The results show that the method based on a search for the shortest version of the name proves to be the easiest, best performing (83.4% F1), and most resource-independent.

Zuo Zhuan Ancient Chinese Dataset for Word Sense Disambiguation
Xiaomeng Pan, Hongfei Wang, Teruaki Oka and Mamoru Komachi 09:15-10:15 (702 Clearwater)
Word Sense Disambiguation (WSD) is a core task in Natural Language Processing (NLP). Ancient Chinese has rarely been used in WSD tasks, however, as no public dataset for ancient Chinese WSD tasks exists. Creation of an ancient Chinese dataset is considered a significant challenge because determining the most appropriate sense in a context is difficult and time-consuming owing to the different usages in ancient and modern Chinese. Actually, no public dataset for ancient Chinese WSD tasks exists. To solve the problem of ancient Chinese WSD, we annotate part of Pre-Qin (221 BC) text extitZuo Zhuan using a copyright-free dictionary to create a public sense-tagged dataset. Then, we apply a simple Nearest Neighbors (k-NN) method using a pre-trained language model to the dataset. Our code and dataset will be available on GitHub
ootnote`https://github.com/pxm427/Ancient-Chinese-WSD`.

CSSS: A Novel Candidate Summary Selection Strategy for Summary-level Extractive Summarization
Shuai Gong, Zhenfang Zhu, Wenqing Wu, Zhen Zhao and Dianyuan Zhang 09:15-10:15 (702 Clearwater)
Summary-level extractive summarization selects a summary with the highest semantic similarity to the document through a matching model, resulting in insufficient use of information between different candidate summaries. This paper presents a novel candidate summary selection strategy (CSSS), regarding candidate summaries as mathematical sets, and selecting the candidate summary has the highest semantic similarity to corresponding mutually exclusive sets. The strategy reduces the dependence on the matching model by exploiting the set relationship, could be effectively applied to both unsupervised and supervised extractive summarization. In order to fit this strategy better, we construct a contrastive learning framework to learn effective vector representation for each candidate summary. Experimental results show that we achieve state-of-the-art performance in both the unsupervised and supervised extractive summarization on CNN/DailyMail dataset. Experiments on Xsum and Reddit datasets also show the effectiveness of CSSS.

Few-shot fine-tuning SOTA summarization models for medical dialogues
David Fraile Navarro, Mark Dras and Shlomo Berkovsky 09:15-10:15 (702 Clearwater)
Abstractive summarization of medical dialogues presents a challenge for standard training approaches, given the paucity of suitable datasets. We explore the performance of state-of-the-art models with zero-shot and few-shot learning strategies and measure the impact of pretraining with general domain and dialogue-specific text on the summarization performance.

Session 9 - 10:45-12:15

Language Generation

10:45-12:00 (Columbia A)

FRUIT: Faithfully Reflecting Updated Information in Text
Robert L. Logan IV, Alexandre Tachard Passos, Sameer Singh and Ming-Wei Chang 10:45-11:00 (Columbia A)
Textual knowledge bases such as Wikipedia require considerable effort to keep up to date and consistent. While automated writing assistants could potentially ease this burden, the problem of suggesting edits grounded in external knowledge has been under-explored. In this paper, we introduce the novel generation task of *faithfully reflecting updated information in text* (FRUIT) where the goal is to update an existing article given new evidence. We release the FRUIT-WIKI dataset, a collection of over 170K distantly supervised data produced from pairs of Wikipedia snapshots, along with our data generation pipeline and a gold evaluation set of 914 instances whose edits are guaranteed to be supported by the evidence. We provide benchmark results for popular generation systems as well as EDIT5 – a T5-based approach tailored to editing we introduce that establishes the state of the art. Our analysis shows that developing models that can update articles faithfully requires new capabilities for neural generation models, and opens doors to many new applications.

Persona-Guided Planning for Controlling the Protagonist's Persona in Story Generation
Zhexin Zhang, Jiaxin Wen, Jian Guan and Minlie Huang 11:00-11:15 (Columbia A)
Endowing the protagonist with a specific personality is essential for writing an engaging story. In this paper, we aim to control the protagonist's persona in story generation, i.e., generating a story from a leading context and a persona description, where the protagonist should exhibit the specified personality through a coherent event sequence. Considering that personas are usually embodied implicitly and sparsely in stories, we propose a planning-based generation model named ConPer to explicitly model the relationship between personas and events. ConPer first plans events of the protagonist's behavior which are motivated by the specified persona through predicting one target sentence, then plans the plot as a sequence of keywords with the guidance of the predicted persona-related events and commonsense knowledge, and finally generates the whole story. Both automatic and manual evaluation results demonstrate that ConPer outperforms state-of-the-art baselines for generating more coherent and persona-controllable stories. Our code is available at `https://github.com/thu-coai/ConPer`.

Fuse It More Deeply! A Variational Transformer with Layer-Wise Latent Variable Inference for Text Generation
Jinyi Hu, Xiaoyuan Yi, Wenhao Li, Maosong Sun and Xing Xie 11:15-11:30 (Columbia A)
The past several years have witnessed Variational Auto-Encoder's superiority in various text generation tasks. However, due to the sequential nature of the text, auto-regressive decoders tend to ignore latent variables and then reduce to simple language models, known as the

KL vanishing problem, which would further deteriorate when VAE is combined with Transformer-based structures. To ameliorate this problem, we propose Della, a novel variational Transformer framework. Della learns a series of layer-wise latent variables with each inferred from those of lower layers and tightly coupled with the hidden states by low-rank tensor product. In this way, Della forces these posterior latent variables to be fused deeply with the whole computation path and hence incorporate more information. We theoretically demonstrate that our method can be regarded as entangling latent variables to avoid posterior information decrease through layers, enabling Della to get higher non-zero KL values even without any annealing or thresholding tricks. Experiments on four unconditional and three conditional generation tasks show that Della could better alleviate KL vanishing and improve both quality and diversity compared to several strong baselines.

Overcoming Catastrophic Forgetting During Domain Adaptation of Seq2seq Language Generation
Dingcheng Li, Zheng Chen, Eunah Cho, Jie Hao, Xiaohu Liu, Fan Xing, Chenlei Guo and Yang Liu 11:30-11:45 (Columbia A)
Seq2seq language generation models that are trained offline with multiple domains in a sequential fashion often suffer from catastrophic forgetting. Lifelong learning has been proposed to handle this problem. However, existing work such as experience replay or elastic weighted consolidation requires incremental memory space. In this work, we propose an innovative framework, RMR_DSEthat leverages a recall optimization mechanism to selectively memorize important parameters of previous tasks via regularization, and uses a domain drift estimation algorithm to compensate the drift between different do-mains in the embedding space. These designs enable the model to be trained on the current task while keep-ing the memory of previous tasks, and avoid much additional data storage. Furthermore, RMR_DSE can be combined with existing lifelong learning approaches. Our experiments on two seq2seq language generation tasks, paraphrase and dialog response generation, show thatRMR_DSE outperforms SOTA models by a considerable margin and reduces forgetting greatly.

Zero-shot Sonnet Generation with Discourse-level Planning and Aesthetics Features
Yufei Tian and Nanyun Peng 11:45-12:00 (Columbia A)
Poetry generation, and creative language generation in general, usually suffers from the lack of large training data. In this paper, we present a novel framework to generate sonnets that does not require training on poems. We design a hierarchical framework which plans the poem sketch before decoding. Specifically, a content planning module is trained on non-poetic texts to obtain discourse-level coherence; then a rhyme module generates rhyme words and a polishing module introduces imagery and similes for aesthetics purposes. Finally, we design a constrained decoding algorithm to impose the meter-and-rhyme constraint of the generated sonnets. Automatic and human evaluation show that our multi-stage approach without training on poem corpora generates more coherent, poetic, and creative sonnets than several strong baselines.

Speech & Phonology, Morphology

10:45-12:00 (Columbia C)

Textless Speech-to-Speech Translation on Real Data
Ann Lee, Hongyu Gong, Paul-Ambroise Duquenne, Holger Schwenk, Peng-Jen Chen, Changhan Wang, Sravya Popuri, Yossi Adi, Juan Pino, Jiatao Gu and Wei-Ning Hsu 10:45-11:00 (Columbia C)
We present a textless speech-to-speech translation (S2ST) system that can translate speech from one language into another language and can be built without the need of any text data. Different from existing work in the literature, we tackle the challenge in modeling multi-speaker target speech and train the systems with real-world S2ST data. The key to our approach is a self-supervised unit-based speech normalization technique, which finetunes a pre-trained speech encoder with paired audios from multiple speakers and a single reference speaker to reduce the variations due to accents, while preserving the lexical content. With only 10 minutes of paired data for speech normalization, we obtain on average 3.2 BLEU gain when training the S2ST model on the VoxPopuli S2ST dataset, compared to a baseline trained on un-normalized speech target. We also incorporate automatically mined S2ST data and show an additional 2.0 BLEU gain. To our knowledge, we are the first to establish a textless S2ST technique that can be trained with real-world data and works for multiple language pairs.

Unsupervised Stem-based Cross-lingual Part-of-Speech Tagging for Morphologically Rich Low-Resource Languages
Ramy Eskander, Cass Lowry, Sujay Khandagale, Judith Lynn Klavans, Maria Polinsky and Smaranda Muresan 11:00-11:15 (Columbia C)
Unsupervised cross-lingual projection for part-of-speech (POS) tagging relies on the use of parallel data to project POS tags from a source language for which a POS tagger is available onto a target language across word-level alignments. The projected tags then form the basis for learning a POS model for the target language. However, languages with rich morphology often yield sparse word alignments because words corresponding to the same citation form do not align well. We hypothesize that for morphologically complex languages, it is more efficient to use the stem rather than the word as the core unit of abstraction. Our contributions are: 1) we propose an unsupervised stem-based cross-lingual approach for POS tagging for low-resource languages of rich morphology; 2) we further investigate morpheme-level alignment and projection; and 3) we examine whether the use of linguistic priors for morphological segmentation improves POS tagging. We conduct experiments using six source languages and eight morphologically complex target languages of diverse typologies. Our results show that the stem-based approach improves the POS models for all the target languages, with an average relative error reduction of 10.3% in accuracy per target language, and outperforms the word-based approach that operates on three-times more data for about two thirds of the language pairs we consider. Moreover, we show that morpheme-level alignment and projection and the use of linguistic priors for morphological segmentation further improve POS tagging.

Quantifying Synthesis and Fusion and their Impact on Machine Translation
Arturo Oncevay, Duygu Ataman, Niels Van Berkel, Barry Haddow, Alexandra Birch and Johannes Bjerva 11:15-11:30 (Columbia C)
Theoretical work in morphological typology offers the possibility of measuring morphological diversity on a continuous scale. However, literature in Natural Language Processing (NLP) typically labels a whole language with a strict type of morphology, e.g. fusional or agglutinative. In this work, we propose to reduce the rigidity of such claims, by quantifying morphological typology at the word and segment level. We consider Payne (2017)'s approach to classify morphology using two indices: synthesis (e.g. analytic to polysynthetic) and fusion (agglutinative to fusional). For computing synthesis, we test unsupervised and supervised morphological segmentation methods for English, German and Turkish, whereas for fusion, we propose a semi-automatic method using Spanish as a case study. Then, we analyse the relationship between machine translation quality and the degree of synthesis and fusion at word (nouns and verbs for English-Turkish, and verbs in English-Spanish) and segment level (previous language pairs plus English-German in both directions). We complement the word-level analysis with human evaluation, and overall, we observe a consistent impact of both indexes on machine translation quality.

On the Use of External Data for Spoken Named Entity Recognition
Ankita Pasad, Felix Wu, Suwon Shon, Karen Livescu and Kyu Han 11:30-11:45 (Columbia C)
Spoken language understanding (SLU) tasks involve mapping from speech signals to semantic labels. Given the complexity of such tasks, good performance is expected to require large labeled datasets, which are difficult to collect for each new task and domain. However, recent

advances in self-supervised speech representations have made it feasible to consider learning SLU models with limited labeled data. In this work, we focus on low-resource spoken named entity recognition (NER) and address the question: Beyond self-supervised pre-training, how can we use external speech and/or text data that are not annotated for the task? We consider self-training, knowledge distillation, and transfer learning for end-to-end (E2E) and pipeline (speech recognition followed by text NER) approaches. We find that several of these approaches improve performance in resource-constrained settings beyond the benefits from pre-trained representations. Compared to prior work, we find relative improvements in F1 of up to 16%. While the best baseline model is a pipeline approach, the best performance using external data is ultimately achieved by an E2E model. We provide detailed comparisons and analyses, developing insights on, for example, the effects of leveraging external data on (i) different categories of NER errors and (ii) the switch in performance trends between pipeline and E2E models.

Empathic Machines: Using Intermediate Features as Levers to Emulate Emotions in Text-To-Speech Systems
Saiteja Kosgi, Sarath Sivaprasad, Niranjan Pedanekar, Anil Kumar Nelakanti and Vineet Gandhi 11:45-12:00 (Columbia C)
We present a method to control the emotional prosody of Text to Speech (TTS) systems by using phoneme-level intermediate features (pitch, energy, and duration) as levers. As a key idea, we propose Differential Scaling (DS) to disentangle features relating to affective prosody from those arising due to acoustics conditions and speaker identity. With thorough experimental studies, we show that the proposed method improves over the prior art in accurately emulating the desired emotions while retaining the naturalness of speech. We extend the traditional evaluation of using individual sentences for a more complete evaluation of HCI systems. We present a novel experimental setup by replacing an actor with a TTS system in offline and live conversations. The emotion to be rendered is either predicted or manually assigned. The results show that the proposed method is strongly preferred over the state-of-the-art TTS system and adds the much-coveted "human touch" in machine dialogue. Audio samples from our experiments and the code are available at: https://emtts.github.io/tts-demo/

Information Extraction 3

10:45-12:15 (Columbia D)

Improving Entity Disambiguation by Reasoning over a Knowledge Base
Tom Ayoola, Joseph Fisher and Andrea Pierleoni 10:45-11:00 (Columbia D)
Recent work in entity disambiguation (ED) has typically neglected structured knowledge base (KB) facts, and instead relied on a limited subset of KB information, such as entity descriptions or types. This limits the range of contexts in which entities can be disambiguated. To allow the use of all KB facts, as well as descriptions and types, we introduce an ED model which links entities by reasoning over a symbolic knowledge base in a fully differentiable fashion. Our model surpasses state-of-the-art baselines on six well-established ED datasets by 1.3 F1 on average. By allowing access to all KB information, our model is less reliant on popularity-based entity priors, and improves performance on the challenging ShadowLink dataset (which emphasises infrequent and ambiguous entities) by 12.7 F1.

DocTime: A Document-level Temporal Dependency Graph Parser
Puneet Mathur, Vlad I Morariu, Verena Kaynig-Fittkau, Jiuxiang Gu, Franck Dernoncourt, Quan Hung Tran, Ani Nenkova, Dinesh Manocha and Rajiv Jain 11:00-11:15 (Columbia D)
We introduce DocTime - a novel temporal dependency graph (TDG) parser that takes as input a text document and produces a temporal dependency graph. It outperforms previous BERT-based solutions by a relative 4-8% on three datasets from modeling the problem as a graph network with path-prediction loss to incorporate longer range dependencies. This work also demonstrates how the TDG graph can be used to improve the downstream tasks of temporal questions answering and NLI by a relative 4-10% with a new framework that incorporates the temporal dependency graph into the self-attention layer of Transformer models (`Time-transformer`). Finally, we develop and evaluate on a new temporal dependency graph dataset for the domain of contractual documents, which has not been previously explored in this setting.

SAIS: Supervising and Augmenting Intermediate Steps for Document-Level Relation Extraction
Yuxin Xiao, Zecheng Zhang, Yuning Mao, Carl Yang and Jiawei Han 11:15-11:30 (Columbia D)
Stepping from sentence-level to document-level, the research on relation extraction (RE) confronts increasing text length and more complicated entity interactions. Consequently, it is more challenging to encode the key information sources—relevant contexts and entity types. However, existing methods only implicitly learn to model these critical information sources while being trained for RE. As a result, they suffer the problems of ineffective supervision and uninterpretable model predictions. In contrast, we propose to explicitly teach the model to capture relevant contexts and entity types by supervising and augmenting intermediate steps (SAIS) for RE. Based on a broad spectrum of carefully designed tasks, our proposed SAIS method not only extracts relations of better quality due to more effective supervision, but also retrieves the corresponding supporting evidence more accurately so as to enhance interpretability. By assessing model uncertainty, SAIS further boosts the performance via evidence-based data augmentation and ensemble inference while reducing the computational cost. Eventually, SAIS delivers state-of-the-art RE results on three benchmarks (DocRED, CDR, and GDA) and outperforms the runner-up by 5.04% relatively in F1 score in evidence retrieval on DocRED.

Should We Rely on Entity Mentions for Relation Extraction? Debiasing Relation Extraction with Counterfactual Analysis
*Yiwei Wang, Muhao Chen, Wenxuan Zhou, Yujun Cai, Yuxuan Liang, Dayiheng Liu, Baosong Yang, Juncheng Liu and Bryan Hooi*11:30-11:45 (Columbia D)
Recent literature focuses on utilizing the entity information in the sentence-level relation extraction (RE), but this risks leaking superficial and spurious clues of relations. As a result, RE still suffers from unintended entity bias, i.e., the spurious correlation between entity mentions (names) and relations. Entity bias can mislead the RE models to extract the relations that do not exist in the text. To combat this issue, some previous work masks the entity mentions to prevent the RE models from over-fitting entity mentions. However, this strategy degrades the RE performance because it loses the semantic information of entities. In this paper, we propose the CoRE (Counterfactual Analysis based Relation Extraction) debiasing method that guides the RE models to focus on the main effects of textual context without losing the entity information. We first construct a causal graph for RE, which models the dependencies between variables in RE models. Then, we propose to conduct counterfactual analysis on our causal graph to distill and mitigate the entity bias, that captures the causal effects of specific entity mentions in each instance. Note that our CoRE method is model-agnostic to debias existing RE systems during inference without changing their training processes. Extensive experimental results demonstrate that our CoRE yields significant gains on both effectiveness and generalization for RE. The source code is provided at: https://github.com/vanoracai/CoRE.

MINION: a Large-Scale and Diverse Dataset for Multilingual Event Detection
Amir Pouran Ben Veyseh, Minh Van Nguyen, Franck Dernoncourt and Thien Huu Nguyen 11:45-12:00 (Columbia D)
Event Detection (ED) is the task of identifying and classifying trigger words of event mentions in text. Despite considerable research efforts in recent years for English text, the task of ED in other languages has been significantly less explored. Switching to non-English languages,

important research questions for ED include how well existing ED models perform on different languages, how challenging ED is in other languages, and how well ED knowledge and annotation can be transferred across languages. To answer those questions, it is crucial to obtain multilingual ED datasets that provide consistent event annotation for multiple languages. There exist some multilingual ED datasets; however, they tend to cover a handful of languages and mainly focus on popular ones. Many languages are not covered in existing multilingual ED datasets. In addition, the current datasets are often small and not accessible to the public. To overcome those shortcomings, we introduce a new large-scale multilingual dataset for ED (called MINION) that consistently annotates events for 8 different languages; 5 of them have not been supported by existing multilingual datasets. We also perform extensive experiments and analysis to demonstrate the challenges and transferability of ED across languages in MINION that in all call for more research effort in this area. We will release the dataset to promote future research on multilingual ED.

GenIE: Generative Information Extraction
Martin Josifoski, Nicola De Cao, Maxime Peyrard, Fabio Petroni and Robert West 12:00-12:15 (Columbia D)
Structured and grounded representation of text is typically formalized by closed information extraction, the problem of extracting an exhaustive set of (subject, relation, object) triplets that are consistent with a predefined set of entities and relations from a knowledge base schema. Most existing works are pipelines prone to error accumulation, and all approaches are only applicable to unrealistically small numbers of entities and relations. We introduce GenIE (generative information extraction), the first end-to-end autoregressive formulation of closed information extraction. GenIE naturally exploits the language knowledge from the pre-trained transformer by autoregressively generating relations and entities in textual form. Thanks to a new bi-level constrained generation strategy, only triplets consistent with the predefined knowledge base schema are produced. Our experiments show that GenIE is state-of-the-art on closed information extraction, generalizes from fewer training data points than baselines, and scales to a previously unmanageable number of entities and relations. With this work, closed information extraction becomes practical in realistic scenarios, providing new opportunities for downstream tasks. Finally, this work paves the way towards a unified end-to-end approach to the core tasks of information extraction.

Language Resources & Evaluation 3

10:45-12:15 (Elwha A)

Models in the Loop: Aiding Crowdworkers with Generative Annotation Assistants
Max Bartolo, Tristan Thrush, Sebastian Riedel, Pontus Stenetorp, Robin Jia and Douwe Kiela 10:45-11:00 (Elwha A)
In Dynamic Adversarial Data Collection (DADC), human annotators are tasked with finding examples that models struggle to predict correctly. Models trained on DADC-collected training data have been shown to be more robust in adversarial and out-of-domain settings, and are considerably harder for humans to fool. However, DADC is more time-consuming than traditional data collection and thus more costly per annotated example. In this work, we examine whether we can maintain the advantages of DADC, without incurring the additional cost. To that end, we introduce Generative Annotation Assistants (GAAs), generator-in-the-loop models that provide real-time suggestions that annotators can either approve, modify, or reject entirely. We collect training datasets in twenty experimental settings and perform a detailed analysis of this approach for the task of extractive question answering (QA) for both standard and adversarial data collection. We demonstrate that GAAs provide significant efficiency benefits with over a 30% annotation speed-up, while leading to over a 5x improvement in model fooling rates. In addition, we find that using GAA-assisted training data leads to higher downstream model performance on a variety of question answering tasks over adversarial data collection.

MultiCite: Modeling realistic citations requires moving beyond the single-sentence single-label setting
Anne Lauscher, Brandon Ko, Bailey Kuehl, Sophie Johnson, Arman Cohan, David Jurgens and Kyle Lo 11:00-11:15 (Elwha A)
Citation context analysis (CCA) is an important task in natural language processing that studies how and why scholars discuss each others' work. Despite decades of study, computational methods for CCA have largely relied on overly-simplistic assumptions of how authors cite, which ignore several important phenomena. For instance, scholarly papers often contain rich discussions of cited work that span multiple sentences and express multiple intents concurrently. Yet, recent work in CCA is often approached as a single-sentence, single-label classification task, and thus many datasets used to develop modern computational approaches fail to capture this interesting discourse. To address this research gap, we highlight three understudied phenomena for CCA and release MULTICITE, a new dataset of 12.6K citation contexts from 1.2K computational linguistics papers that fully models these phenomena. Not only is it the largest collection of expert-annotated citation contexts to-date, MULTICITE contains multi-sentence, multi-label citation contexts annotated through-out entire full paper texts. We demonstrate how MULTICITE can enable the development of new computational methods on three important CCA tasks. We release our code and dataset at https://github.com/allenai/multicite.

NewsEdits: A News Article Revision Dataset and a Novel Document-Level Reasoning Challenge
Alexander Spangher, Xiang Ren, Jonathan May and Nanyun Peng 11:15-11:30 (Elwha A)
News article revision histories provide clues to narrative and factual evolution in news articles. To facilitate analysis of this evolution, we present the first publicly available dataset of news revision histories, NewsEdits. Our dataset is large-scale and multilingual; it contains 1.2 million articles with 4.6 million versions from over 22 English- and French-language newspaper sources based in three countries, spanning 15 years of coverage (2006-2021).
We define article-level edit actions: Addition, Deletion, Edit and Refactor, and develop a high-accuracy extraction algorithm to identify these actions. To underscore the factual nature of many edit actions, we conduct analyses showing that added and deleted sentences are more likely to contain updating events, main content and quotes than unchanged sentences.
Finally, to explore whether edit actions are predictable, we introduce three novel tasks aimed at predicting actions performed during version updates. We show that these tasks are possible for expert humans but are challenging for large NLP models. We hope this can spur research in narrative framing and help provide predictive tools for journalists chasing breaking news.

Explaining Dialogue Evaluation Metrics using Adversarial Behavioral Analysis
Baber Khalid and Sungjin Lee 11:30-11:45 (Elwha A)
There is an increasing trend in using neural methods for dialogue model evaluation. Lack of a framework to investigate these metrics can cause dialogue models to reflect their biases and cause unforeseen problems during interactions. In this work, we propose an adversarial test-suite which generates problematic variations of various dialogue aspects, e.g. logical entailment, using automatic heuristics. We show that dialogue metrics for both open-domain and task-oriented settings are biased in their assessments of different conversation behaviors and fail to properly penalize problematic conversations, by analyzing their assessments of these problematic examples. We conclude that variability in training methodologies and data-induced biases are some of the main causes of these problems. We also conduct an investigation into the metric behaviors using a black-box interpretability model which corroborates our findings and provides evidence that metrics pay attention to the problematic conversational constructs signaling a misunderstanding of different conversation semantics.

Answer Consolidation: Formulation and Benchmarking
Wenxuan Zhou, Qiang Ning, Heba Elfardy, Kevin Small and Muhao Chen 11:45-12:00 (Elwha A)
Current question answering (QA) systems primarily consider the single-answer scenario, where each question is assumed to be paired with one correct answer. However, in many real-world QA applications, multiple answer scenarios arise where consolidating answers into a comprehensive and non-redundant set of answers is a more efficient user interface. In this paper, we formulate the problem of answer consolidation, where answers are partitioned into multiple groups, each representing different aspects of the answer set. Then, given this partitioning, a comprehensive and non-redundant set of answers can be constructed by picking one answer from each group. To initiate research on answer consolidation, we construct a dataset consisting of 4,699 questions and 24,006 sentences and evaluate multiple models. Despite a promising performance achieved by the best-performing supervised models, we still believe this task has room for further improvements.

[TACL] **Data-driven Model Generalizability in Crosslinguistic Low-resource Morphological Segmentation**
Zoey Liu and Emily Prud'hommeau 12:00-12:15 (Elwha A)
Common designs of model evaluation typically focus on monolingual settings, where different models are compared according to their performance on a single data set that is assumed to be representative of all possible data for the task at hand. While this may be reasonable for a large data set, this assumption is difficult to maintain in low-resource scenarios, where artifacts of the data collection can yield data sets that are outliers, potentially making conclusions about model performance coincidental. To address these concerns, we investigate model generalizability in crosslinguistic low-resource scenarios. Using morphological segmentation as the test case, we compare three broad classes of models with different parameterizations, taking data from 11 languages across 6 language families. In each experimental setting, we evaluate all models on a first data set , then examine their performance consistency when introducing new randomly sampled data sets with the same size and when applying the trained models to unseen test sets of varying sizes. The results demonstrate that the extent of model generalization depends on the characteristics of the data set, and does not necessarily rely heavily on the data set size. Among the characteristics that we studied, the ratio of morpheme overlap and that of the average number of morphemes per word between the training and test sets are the two most prominent factors. Our findings suggest that future work should adopt random sampling to construct data sets with different sizes in order to make more responsible claims about model evaluation.

NLP Applications 4

10:45-12:15 (Elwha B)

Efficient Constituency Tree based Encoding for Natural Language to Bash Translation
Shikhar Bharadwaj and Shirish Shevade 10:45-11:00 (Elwha B)
Bash is a Unix command language used for interacting with the Operating System. Recent works on natural language to Bash translation have made significant advances, but none of the previous methods utilize the problem's inherent structure. We identify this structure and propose a Segmented Invocation Transformer (SIT) that utilizes the information from the constituency parse tree of the natural language text. Our method is motivated by the alignment between segments in the natural language text and Bash command components. Incorporating the structure in the modelling improves the performance of the model. Since such systems must be universally accessible, we benchmark the inference times on a CPU rather than a GPU. We observe a 1.8x improvement in the inference time and a 5x reduction in model parameters. Attribution analysis using Integrated Gradients reveals that the proposed method can capture the problem structure.

Multi-Relational Graph Transformer for Automatic Short Answer Grading
Rajat Agarwal, Varun Khurana, Karish Grover, Mukesh Mohania and Vikram Goyal 11:00-11:15 (Elwha B)
The recent transition to the online educational domain has increased the need for Automatic Short Answer Grading (ASAG). ASAG automatically evaluates a student's response against a (given) correct response and thus has been a prevalent semantic matching task. Most existing methods utilize sequential context to compare two sentences and ignore the structural context of the sentence; therefore, these methods may not result in the desired performance. In this paper, we overcome this problem by proposing a Multi-Relational Graph Transformer, MitiGaTe, to prepare token representations considering the structural context. Abstract Meaning Representation (AMR) graph is created by parsing the text response and then segregated into multiple subgraphs, each corresponding to a particular relationship in AMR. A Graph Transformer is used to prepare relation-specific token embeddings within each subgraph, then aggregated to obtain a subgraph representation. Finally, we compare the correct answer and the student response subgraph representations to yield a final score. Experimental results on Mohler's dataset show that our system outperforms the existing state-of-the-art methods. We have released our implementation https://github.com/kvarun07/asag-gt, as we believe that our model can be useful for many future applications.

ConfliBERT: A Pre-trained Language Model for Political Conflict and Violence
Yibo Hu, MohammadSaleh Hosseini, Erick Skorupa Parolin, Javier Osorio, Latifur Khan, Patrick Brandt and Vito D'Orazio 11:15-11:30 (Elwha B)
Analyzing conflicts and political violence around the world is a persistent challenge in the political science and policy communities due in large part to the vast volumes of specialized text needed to monitor conflict and violence on a global scale. To help advance research in political science, we introduce ConfliBERT, a domain-specific pre-trained language model for conflict and political violence. We first gather a large domain-specific text corpus for language modeling from various sources. We then build ConfliBERT using two approaches: pre-training from scratch and continual pre-training. To evaluate ConfliBERT, we collect 12 datasets and implement 18 tasks to assess the models' practical application in conflict research. Finally, we evaluate several versions of ConfliBERT in multiple experiments. Results consistently show that ConfliBERT outperforms BERT when analyzing political violence and conflict.

Semantically Informed Slang Interpretation
Zhewei Sun, Richard Zemel and Yang Xu 11:30-11:45 (Elwha B)
Slang is a predominant form of informal language making flexible and extended use of words that is notoriously hard for natural language processing systems to interpret. Existing approaches to slang interpretation tend to rely on context but ignore semantic extensions common in slang word usage. We propose a semantically informed slang interpretation (SSI) framework that considers jointly the contextual and semantic appropriateness of a candidate interpretation for a query slang. We perform rigorous evaluation on two large-scale online slang dictionaries and show that our approach not only achieves state-of-the-art accuracy for slang interpretation in English, but also does so in zero-shot and few-shot scenarios where training data is sparse. Furthermore, we show how the same framework can be applied to enhancing machine translation of slang from English to other languages. Our work creates opportunities for the automated interpretation and translation of informal language.

Don't sweat the small stuff, classify the rest: Sample Shielding to protect text classifiers against adversarial attacks

Jonathan Rusert and Padmini Srinivasan 11:45-12:00 (Elwha B)
Deep learning (DL) is being used extensively for text classification. However, researchers have demonstrated the vulnerability of such classifiers to adversarial attacks. Attackers modify the text in a way which misleads the classifier while keeping the original meaning close to intact. State-of-the-art (SOTA) attack algorithms follow the general principle of making minimal changes to the text so as to not jeopardize semantics. Taking advantage of this we propose a novel and intuitive defense strategy called Sample Shielding. It is attacker and classifier agnostic, does not require any reconfiguration of the classifier or external resources and is simple to implement. Essentially, we sample subsets of the input text, classify them and summarize these into a final decision. We shield three popular DL text classifiers with Sample Shielding, test their resilience against four SOTA attackers across three datasets in a realistic threat setting. Even when given the advantage of knowing about our shielding strategy the adversary's attack success rate is <=10% with only one exception and often < 5%. Additionally, Sample Shielding maintains near original accuracy when applied to original texts. Crucially, we show that the 'make minimal changes' approach of SOTA attackers leads to critical vulnerabilities that can be defended against with an intuitive sampling strategy.

GRAM: Fast Fine-tuning of Pre-trained Language Models for Content-based Collaborative Filtering
Yoonseok Yang, Kyu Seok Kim, Minsam Kim and Juneyoung Park 12:00-12:15 (Elwha B)
Content-based collaborative filtering (CCF) predicts user-item interactions based on both users' interaction history and items' content information. Recently, pre-trained language models (PLM) have been used to extract high-quality item encodings for CCF. However, it is resource-intensive to train a PLM-based CCF model in an end-to-end (E2E) manner, since optimization involves back-propagating through every content encoding within a given user interaction sequence. To tackle this issue, we propose GRAM (GRadient Accumulation for Multi-modality in CCF), which exploits the fact that a given item often appears multiple times within a batch of interaction histories. Specifically, Single-step GRAM aggregates each item encoding's gradients for back-propagation, with theoretic equivalence to the standard E2E training. As an extension of Single-step GRAM, we propose Multi-step GRAM, which increases the gradient update latency, achieving a further speedup with drastically less GPU memory. GRAM significantly improves training efficiency (up to 146x) on five datasets from two task domains of Knowledge Tracing and News Recommendation. Our code is available at https://github.com/yoonseok312/GRAM.

Findings Poster Session 1

10:45-12:15 (Regency A & B)

Modeling Ideological Salience and Framing in Polarized Online Groups with Graph Neural Networks and Structured Sparsity
Valentin Hofmann, Xiaowen Dong, Janet B. Pierrehumbert and Hinrich Schuetze 10:45-12:15 (Regency A & B)
The increasing polarization of online political discourse calls for computational tools that automatically detect and monitor ideological divides in social media. We introduce a minimally supervised method that leverages the network structure of online discussion forums, specifically Reddit, to detect polarized concepts. We model polarization along the dimensions of salience and framing, drawing upon insights from moral psychology. Our architecture combines graph neural networks with structured sparsity learning and results in representations for concepts and subreddits that capture temporal ideological dynamics such as right-wing and left-wing radicalization.

HUE: Pretrained Model and Dataset for Understanding Hanja Documents of Ancient Korea
Haneul Yoo, Jiho Jin, Juhee Son, JinYeong Bak, Kyunghyun Cho and Alice Oh 10:45-12:15 (Regency A & B)
Historical records in Korea before the 20th century were primarily written in Hanja, an extinct language based on Chinese characters and not understood by modern Korean or Chinese speakers. Historians with expertise in this time period have been analyzing the documents, but that process is very difficult and time-consuming, and language models would significantly speed up the process. Toward building and evaluating language models for Hanja, we release the Hanja Understanding Evaluation dataset consisting of chronological attribution, topic classification, named entity recognition, and summary retrieval tasks. We also present BERT-based models continued training on the two major corpora from the 14th to the 19th centuries: the Annals of the Joseon Dynasty and Diaries of the Royal Secretariats. We compare the models with several baselines on all tasks and show there are significant improvements gained by training on the two corpora. Additionally, we run zero-shot experiments on the Daily Records of the Royal Court and Important Officials (DRRI). The DRRI dataset has not been studied much by the historians, and not at all by the NLP community.

Empathetic Persuasion: Reinforcing Empathy and Persuasiveness in Dialogue Systems
Azlaan Mustafa Samad, Kshitij Mishra, Mauajama Firdaus and Asif Ekbal 10:45-12:15 (Regency A & B)
Persuasion is an intricate process involving empathetic connection between two individuals. Plain persuasive responses may make a conversation non-engaging. Even the most well-intended and reasoned persuasive conversations can fall through in the absence of empathetic connection between the speaker and listener. In this paper, we propose a novel task of incorporating empathy when generating persuasive responses. We develop an empathetic persuasive dialogue system by fine-tuning a maximum likelihood Estimation (MLE)-based language model in a reinforcement learning (RL) framework. To design feedbacks for our RL-agent, we define an effective and efficient reward function considering consistency, repetitiveness, emotion and persuasion rewards to ensure consistency, non-repetitiveness, empathy and persuasiveness in the generated responses. Due to lack of emotion annotated persuasive data, we first annotate the existing Persuaion For Good dataset with emotions, then build transformer based classifiers to provide emotion based feedbacks to our RL agent. Experimental results confirm that our proposed model increases the rate of generating persuasive responses as compared to the available state-of-the-art dialogue models while making the dialogues empathetically more engaging and retaining the language quality in responses.

Target-Guided Dialogue Response Generation Using Commonsense and Data Augmentation
Prakhar Gupta, Harsh Jhamtani and Jeffrey Bigham 10:45-12:15 (Regency A & B)
Target-guided response generation enables dialogue systems to smoothly transition a conversation from a dialogue context toward a target sentence. Such control is useful for designing dialogue systems that direct a conversation toward specific goals, such as creating non-obtrusive recommendations or introducing new topics in the conversation. In this paper, we introduce a new technique for target-guided response generation, which first finds a bridging path of commonsense knowledge concepts between the source and the target, and then uses the identified bridging path to generate transition responses. Additionally, we propose techniques to re-purpose existing dialogue datasets for target-guided generation. Experiments reveal that the proposed techniques outperform various baselines on this task. Finally, we observe that the existing automated metrics for this task correlate poorly with human judgement ratings. We propose a novel evaluation metric that we demonstrate is more reliable for target-guided response evaluation. Our work generally enables dialogue system designers to exercise more control over the conversations that their systems produce.

Balancing Multi-Domain Corpora Learning for Open-Domain Response Generation
Yujie Xing, Jinglun Cai, Nils Barlaug, Peng Liu and Jon Atle Gulla 10:45-12:15 (Regency A & B)
Open-domain conversational systems are assumed to generate equally good responses on multiple domains. Previous work achieved good

performance on the single corpus, but training and evaluating on multiple corpora from different domains are less studied. This paper explores methods of generating relevant responses for each of multiple multi-domain corpora. We first examine interleaved learning which intermingles multiple corpora as the baseline. We then investigate two multi-domain learning methods, labeled learning and multi-task labeled learning, which encode each corpus through a unique corpus embedding. Furthermore, we propose Domain-specific Frequency (DF), a novel word-level importance weight that measures the relative importance of a word for a specific corpus compared to other corpora. Based on DF, we propose weighted learning, a method that integrates DF to the loss function. We also adopt DF as a new evaluation metric. Extensive experiments show that our methods gain significant improvements on both automatic and human evaluation. We share our code and data for reproducibility.

Context-Aware Language Modeling for Goal-Oriented Dialogue Systems
Charlie Victor Snell, Sherry Yang, Justin Fu, Yi Su and Sergey Levine 10:45-12:15 (Regency A & B)
Goal-oriented dialogue systems face a trade-off between fluent language generation and task-specific control. While supervised learning with large language models is capable of producing realistic text, how to steer such responses towards completing a specific task without sacrificing language quality remains an open question. In this work, we formulate goal-oriented dialogue as a partially observed Markov decision process, interpreting the language model as a representation of both the dynamics and the policy. This view allows us to extend techniques from learning-based control, such as task relabeling, to derive a simple and effective method to finetune language models in a goal-aware way, leading to significantly improved task performance. We additionally introduce a number of training strategies that serve to better focus the model on the task at hand. We evaluate our method, Context-Aware Language Models (CALM), on a practical flight-booking task using AirDialogue. Empirically, CALM outperforms the state-of-the-art method by 7% in terms of task success, matching human-level task performance.

KETOD: Knowledge-Enriched Task-Oriented Dialogue
Zhiyu Chen, Bing Liu, Seungwhan Moon, Chinnadhurai Sankar, Paul A. Crook and William Yang Wang 10:45-12:15 (Regency A & B)
Existing studies in dialogue system research mostly treat task-oriented dialogue and chit-chat as separate domains. Towards building a human-like assistant that can converse naturally and seamlessly with users, it is important to build a dialogue system that conducts both types of conversations effectively. In this work, we investigate how task-oriented dialogue and knowledge-grounded chit-chat can be effectively integrated into a single model. To this end, we create a new dataset, KETOD (Knowledge-Enriched Task-Oriented Dialogue), where we naturally enrich task-oriented dialogues with chit-chat based on relevant entity knowledge. We also propose two new models, SimpleToD-Plus and Combiner, for the proposed task. Experimental results on both automatic and human evaluations show that the proposed methods can significantly improve the performance in knowledge-enriched response generation while maintaining a competitive task-oriented dialog performance. We believe our new dataset will be a valuable resource for future studies. Our dataset and code are publicly available at `https://github.com/facebookresearch/ketod`.

Improve Discourse Dependency Parsing with Contextualized Representations
Yifei Zhou and Yansong Feng 10:45-12:15 (Regency A & B)
Previous works show that discourse analysis benefits from modeling intra- and inter-sentential levels separately, where proper representations for text units of different granularities are desired to capture both the information of the text units and their relation to the context. In this paper, we propose to take advantage of transformers to encode different contextualized representations of units of different levels to dynamically capture the information required for discourse dependency analysis on intra- and inter-sentential levels. Motivated by the observation of writing patterns shared across articles to improve discourse analysis, we propose to design sequence labeling methods to take advantage of such structural information from the context that substantially outperforms traditional direct classification methods. Experiments show that our model achieves state-of-the-art results on both English and Chinese datasets.

Empowering parameter-efficient transfer learning by recognizing the kernel structure in self-attention
Yifan Chen, Devamanyu Hazarika, Mahdi Namazifar, Yang Liu, Di Jin and Dilek Hakkani-Tur 10:45-12:15 (Regency A & B)
The massive amount of trainable parameters in the pre-trained language models (PLMs) makes them hard to be deployed to multiple downstream tasks. To address this issue, parameter-efficient transfer learning methods have been proposed to tune only a few parameters during fine-tuning while freezing the rest. This paper looks at existing methods along this line through the *kernel lens*. Motivated by the connection between self-attention in transformer-based PLMs and kernel learning, we propose *kernel-wise adapters*, namely *Kernel-mix*, that utilize the kernel structure in self-attention to guide the assignment of the tunable parameters. These adapters use guidelines found in classical kernel learning and enable separate parameter tuning for each attention head. Our empirical results, over a diverse set of natural language generation and understanding tasks, show that our proposed adapters can attain or improve the strong performance of existing baselines.

Great Power, Great Responsibility: Recommendations for Reducing Energy for Training Language Models
Joseph McDonald, Baolin Li, Nathan C. Frey, Devesh Tiwari, Vijay Gadepally and Siddharth Samsi 10:45-12:15 (Regency A & B)
The energy requirements of current natural language processing models continue to grow at a rapid, unsustainable pace. Recent works highlighting this problem conclude there is an urgent need for methods that reduce the energy needs of NLP and machine learning more broadly. In this article, we investigate techniques that can be used to reduce the energy consumption of common NLP applications. In particular, we focus on techniques to measure energy usage and different hardware and datacenter-oriented settings that can be tuned to reduce energy consumption for training and inference for language models. We characterize the impact of these settings on metrics such as computational performance and energy consumption through experiments conducted on a high performance computing system as well as popular cloud computing platforms. These techniques can lead to significant reduction in energy consumption when training language models or their use for inference. For example, power-capping, which limits the maximum power a GPU can consume, can enable a 15% decrease in energy usage with marginal increase in overall computation time when training a transformer-based language model.

AdapterBias: Parameter-efficient Token-dependent Representation Shift for Adapters in NLP Tasks
Chin-Lun Fu, Zih-Ching Chen, Yun-Ru Lee and Hung-yi Lee 10:45-12:15 (Regency A & B)
Transformer-based pre-trained models with millions of parameters require large storage. Recent approaches tackle this shortcoming by training adapters, but these approaches still require a relatively large number of parameters. In this study, AdapterBias, a surprisingly simple yet effective adapter architecture, is proposed. AdapterBias adds a token-dependent shift to the hidden output of transformer layers to adapt to downstream tasks with only a vector and a linear layer. Extensive experiments are conducted to demonstrate the effectiveness of AdapterBias. The experiments show that our proposed method can dramatically reduce the trainable parameters compared to the previous works with a minimal decrease in task performances compared with fine-tuned pre-trained models. We further find that AdapterBias automatically learns to assign more significant representation shifts to the tokens related to the task in consideration.

LongT5: Efficient Text-To-Text Transformer for Long Sequences
Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung and Yinfei Yang 10:45-12:15 (Regency A & B)
Recent work has shown that either (1) increasing the input length or (2) increasing model size can improve the performance of Transformer-based neural models. In this paper, we present LongT5, a new model that explores the effects of scaling both the input length and model size at the same time. Specifically, we integrate attention ideas from long-input transformers (ETC), and adopt pre-training strategies from

summarization pre-training (PEGASUS) into the scalable T5 architecture. The result is a new attention mechanism we call Transient Global (TGlobal), which mimics ETC's local/global attention mechanism, but without requiring additional side-inputs. We are able to achieve state-of-the-art results on several summarization and question answering tasks, as well as outperform the original T5 models on these tasks. We have open sourced our architecture and training code, as well as our pre-trained model checkpoints.

LM-CORE: Language Models with Contextually Relevant External Knowledge
Jivat Neet Kaur, Sumit Bhatia, Milan Aggarwal, Rachit Bansal and Balaji Krishnamurthy 10:45-12:15 (Regency A & B)
Large transformer-based pre-trained language models have achieved impressive performance on a variety of knowledge-intensive tasks and can capture factual knowledge in their parameters. We argue that storing large amounts of knowledge in the model parameters is sub-optimal given the ever-growing amounts of knowledge and resource requirements. We posit that a more efficient alternative is to provide explicit access to contextually relevant structured knowledge to the model and train it to use that knowledge. We present LM-CORE – a general framework to achieve this– that allows *decoupling* of the language model training from the external knowledge source and allows the latter to be updated without affecting the already trained model. Experimental results show that LM-CORE, having access to external knowledge, achieves significant and robust outperformance over state-of-the-art knowledge-enhanced language models on knowledge probing tasks; can effectively handle knowledge updates; and performs well on two downstream tasks. We also present a thorough error analysis highlighting the successes and failures of LM-CORE. Our code and model checkpoints are publicly available.

Cross-Domain Classification of Moral Values
Enrico Liscio, Alin Eugen Dondera, Andrei Geadau, Catholijn M Jonker and Pradeep Kumar Murukannaiah 10:45-12:15 (Regency A & B)
Moral values influence how we interpret and act upon the information we receive. Identifying human moral values is essential for artificially intelligent agents to co-exist with humans. Recent progress in natural language processing allows the identification of moral values in textual discourse. However, domain-specific moral rhetoric poses challenges for transferring knowledge from one domain to another. We provide the first extensive investigation on the effects of cross-domain classification of moral values from text. We compare a state-of-the-art deep learning model (BERT) in seven domains and four cross-domain settings. We show that a value classifier can generalize and transfer knowledge to novel domains, but it can introduce catastrophic forgetting. We also highlight the typical classification errors in cross-domain value classification and compare the model predictions to the annotators agreement. Our results provide insights to computer and social scientists that seek to identify moral rhetoric specific to a domain of discourse.

On Measuring Social Biases in Prompt-Based Multi-Task Learning
Afra Feyza Akyürek, Sejin Paik, Muhammed Yusuf Kocyigit, Seda Akbiyik, Serife Leman Runyun and Derry Wijaya 10:45-12:15 (Regency A & B)
Large language models trained on a mixture of NLP tasks that are converted into a text-to-text format using prompts, can generalize into novel forms of language and handle novel tasks. A large body of work within prompt engineering attempts to understand the effects of input forms and prompts in achieving superior performance. We consider an alternative measure and inquire whether the way in which an input is encoded affects social biases promoted in outputs. In this paper, we study T0, a large-scale multi-task text-to-text language model trained using prompt-based learning. We consider two different forms of semantically equivalent inputs: question-answer format and premise-hypothesis format. We use an existing bias benchmark for the former BBQ and create the first bias benchmark in natural language inference BBNLI with hand-written hypotheses while also converting each benchmark into the other form. The results on two benchmarks suggest that given two different formulations of essentially the same input, T0 conspicuously acts more biased in question answering form, which is seen during training, compared to premise-hypothesis form which is unlike its training examples. Code and data are released under https://github.com/feyzaakyurek/bbnli.

Few-Shot Self-Rationalization with Natural Language Prompts
Ana Marasovic, Iz Beltagy, Doug Downey and Matthew E Peters 10:45-12:15 (Regency A & B)
Self-rationalization models that predict task labels and generate free-text elaborations for their predictions could enable more intuitive interaction with NLP systems. These models are, however, currently trained with a large amount of human-written free-text explanations for each task which hinders their broader usage. We propose to study a more realistic setting of self-rationalization using few training examples. We present FEB—a standardized collection of four existing English-language datasets and associated metrics. We identify the right prompting approach by extensively exploring natural language prompts on FEB. Then, by using this prompt and scaling the model size, we demonstrate that making progress on few-shot self-rationalization is possible. We show there is still ample room for improvement in this task: the average plausibility of generated explanations assessed by human annotators is at most 51% (with GPT-3), while plausibility of human explanations is 76%. We hope that FEB and our proposed approach will spur the community to take on the few-shot self-rationalization challenge.

Entailment Tree Explanations via Iterative Retrieval-Generation Reasoner
Danilo Neves Ribeiro, Shen Wang, Xiaofei Ma, Rui Dong, Xiaokai Wei, Henghui Zhu, Xinchi Chen, Peng Xu, Zhiheng Huang, Andrew Arnold and Dan Roth 10:45-12:15 (Regency A & B)
Large language models have achieved high performance on various question answering (QA) benchmarks, but the explainability of their output remains elusive. Structured explanations, called entailment trees, were recently suggested as a way to explain the reasoning behind a QA system's answer. In order to better generate such entailment trees, we propose an architecture called Iterative Retrieval-Generation Reasoner (IRGR). Our model is able to explain a given hypothesis by systematically generating a step-by-step explanation from textual premises. The IRGR model iteratively searches for suitable premises, constructing a single entailment step at a time. Contrary to previous approaches, our method combines generation steps and retrieval of premises, allowing the model to leverage intermediate conclusions, and mitigating the input size limit of baseline encoder-decoder models. We conduct experiments using the EntailmentBank dataset, where we outperform existing benchmarks on premise retrieval and entailment tree generation, with around 300% gain in overall correctness.

Identifying and Mitigating Spurious Correlations for Improving Robustness in NLP Models
Tianlu Wang, Rohit Sridhar, Diyi Yang and Xuezhi Wang 10:45-12:15 (Regency A & B)
Recently, NLP models have achieved remarkable progress across a variety of tasks; however, they have also been criticized for being not robust. Many robustness problems can be attributed to models exploiting "spurious correlations", or "shortcuts" between the training data and the task labels. Most existing work identifies a limited set of task-specific shortcuts via human priors or error analyses, which requires extensive expertise and efforts. In this paper, we aim to automatically identify such spurious correlations in NLP models at scale. We first leverage existing interpretability methods to extract tokens that significantly affect model's decision process from the input text. We then distinguish "genuine" tokens and "spurious" tokens by analyzing model predictions across multiple corpora and further verify them through knowledge-aware perturbations. We show that our proposed method can effectively and efficiently identify a scalable set of “shortcuts”, and mitigating these leads to more robust models in multiple applications.

Exploring the Universal Vulnerability of Prompt-based Learning Paradigm
Lei Xu, Yangyi Chen, Ganqu Cui, Hongcheng Gao and Zhiyuan Liu 10:45-12:15 (Regency A & B)
Prompt-based learning paradigm bridges the gap between pre-training and fine-tuning, and works effectively under the few-shot setting.

However, we find that this learning paradigm inherits the vulnerability from the pre-training stage, where model predictions can be misled by inserting certain triggers into the text. In this paper, we explore this universal vulnerability by either injecting backdoor triggers or searching for adversarial triggers on pre-trained language models using only plain text. In both scenarios, we demonstrate that our triggers can totally control or severely decrease the performance of prompt-based models fine-tuned on arbitrary downstream tasks, reflecting the universal vulnerability of the prompt-based learning paradigm. Further experiments show that adversarial triggers have good transferability among language models. We also find conventional fine-tuning models are not vulnerable to adversarial triggers constructed from pre-trained language models. We conclude by proposing a potential solution to mitigate our attack methods. Code and data are publicly available.

On the Limitations of Dataset Balancing: The Lost Battle Against Spurious Correlations
Roy Schwartz and Gabriel Stanovsky 10:45-12:15 (Regency A & B)
Recent work has shown that deep learning models in NLP are highly sensitive to low-level correlations between simple features and specific output labels, leading to over-fitting and lack of generalization. To mitigate this problem, a common practice is to balance datasets by adding new instances or by filtering out "easy" instances (Sakaguchi et al., 2020), culminating in a recent proposal to eliminate single-word correlations altogether (Gardner et al., 2021). In this opinion paper, we identify that despite these efforts, increasingly-powerful models keep exploiting ever-smaller spurious correlations, and as a result even balancing all single-word features is insufficient for mitigating all of these correlations. In parallel, a truly balanced dataset may be bound to "throw the baby out with the bathwater" and miss important signal encoding common sense and world knowledge. We highlight several alternatives to dataset balancing, focusing on enhancing datasets with richer contexts, allowing models to abstain and interact with users, and turning from large-scale fine-tuning to zero- or few-shot setups.

Beyond Distributional Hypothesis: Let Language Models Learn Meaning-Text Correspondence
M.j Jang, Frank Martin Mtumbuka and Thomas Lukasiewicz 10:45-12:15 (Regency A & B)
The logical negation property (LNP), which implies generating different predictions for semantically opposite inputs (p is true iff ¬p is false), is an important property that a trustworthy language model must satisfy. However, much recent evidence shows that large-size pre-trained language models (PLMs) do not satisfy this property. In this paper, we perform experiments using probing tasks to assess PLMs' LNP understanding. Unlike previous studies that only examined negation expressions, we expand the boundary of the investigation to lexical semantics. Through experiments, we observe that PLMs violate the LNP frequently. To alleviate the issue, we propose a novel intermediate training task, named meaning-matching, designed to directly learn a meaning text correspondence, instead of relying on the distributional hypothesis. Through multiple experiments, we find that the task enables PLMs to learn lexical semantic information. Also, through fine-tuning experiments on 7 GLUE tasks, we confirm that it is a safe intermediate task that guarantees a similar or better performance of downstream tasks. Finally, we observe that our proposed approach outperforms our previous counterparts despite its time and resource efficiency.

Entity Cloze By Date: What LMs Know About Unseen Entities
Yasumasa Onoe, Michael JQ Zhang, Eunsol Choi and Greg Durrett 10:45-12:15 (Regency A & B)
Language models (LMs) are typically trained once on a large-scale corpus and used for years without being updated. However, in a dynamic world, new entities constantly arise. We propose a framework to analyze what LMs can infer about new entities that did not exist when the LMs were pretrained. We derive a dataset of entities indexed by their origination date and paired with their English Wikipedia articles, from which we can find sentences about each entity. We evaluate LMs' perplexity on masked spans within these sentences. We show that models more informed about the entities, such as those with access to a textual definition of them, achieve lower perplexity on this benchmark. Our experimental results demonstrate that making inferences about new entities remains difficult for LMs. Given its wide coverage on entity knowledge and temporal indexing, our dataset can be used to evaluate LMs and techniques designed to modify or extend their knowledge. Our automatic data collection pipeline can be easily used to continually update our benchmark.

Masked Measurement Prediction: Learning to Jointly Predict Quantities and Units from Textual Context
Daniel Spokoyny, Ivan Lee, Zhao Jin and Taylor Berg-Kirkpatrick 10:45-12:15 (Regency A & B)
Physical measurements constitute a large portion of numbers in academic papers, engineering reports, and web tables. Current benchmarks fall short of properly evaluating numeracy of pretrained language models on measurements, hindering research on developing new methods and applying them to numerical tasks. To that end, we introduce a novel task, Masked Measurement Prediction (MMP), where a model learns to reconstruct a number together with its associated unit given masked text. MMP is useful for both training new numerically informed models as well as evaluating numeracy of existing systems. To address this task, we introduce a new Generative Masked Measurement (GeMM) model that jointly learns to predict numbers along with their units. We perform fine-grained analyses comparing our model with various ablations and baselines. We use linear probing of traditional pretrained transformer models (RoBERTa) to show that they significantly underperform jointly trained number-unit models, highlighting the difficulty of this new task and the benefits of our proposed pretraining approach. We hope this framework accelerates the progress towards building more robust numerical reasoning systems in the future.

Learning Rich Representation of Keyphrases from Text
Mayank Kulkarni, Debanjan Mahata, Ravneet Singh Arora and Rajarshi Bhowmik 10:45-12:15 (Regency A & B)
In this work, we explore how to train task-specific language models aimed towards learning rich representation of keyphrases from text documents. We experiment with different masking strategies for pre-training transformer language models (LMs) in discriminative as well as generative settings. In the discriminative setting, we introduce a new pre-training objective - Keyphrase Boundary Infilling with Replacement (KBIR), showing large gains in performance (upto 8.16 points in F1) over SOTA, when the LM pre-trained using KBIR is fine-tuned for the task of keyphrase extraction. In the generative setting, we introduce a new pre-training setup for BART - KeyBART, that reproduces the keyphrases related to the input text in the CatSeq format, instead of the denoised original input. This also led to gains in performance (upto 4.33 points in F1@M) over SOTA for keyphrase generation. Additionally, we also fine-tune the pre-trained language models on named entity recognition (NER), question answering (QA), relation extraction (RE), abstractive summarization and achieve comparable performance with that of the SOTA, showing that learning rich representation of keyphrases is indeed beneficial for many other fundamental NLP tasks.

Temporal Attention for Language Models
Guy D. Rosin and Kira Radinsky 10:45-12:15 (Regency A & B)
Pretrained language models based on the transformer architecture have shown great success in NLP. Textual training data often comes from the web and is thus tagged with time-specific information, but most language models ignore this information. They are trained on the textual data alone, limiting their ability to generalize temporally. In this work, we extend the key component of the transformer architecture, i.e., the self-attention mechanism, and propose temporal attention - a time-aware self-attention mechanism. Temporal attention can be applied to any transformer model and requires the input texts to be accompanied with their relevant time points. This mechanism allows the transformer to capture this temporal information and create time-specific contextualized word representations. We leverage these representations for the task of semantic change detection; we apply our proposed mechanism to BERT and experiment on three datasets in different languages (English, German, and Latin) that also vary in time, size, and genre. Our proposed model achieves state-of-the-art results on all the datasets.

Lacuna Reconstruction: Self-Supervised Pre-Training for Low-Resource Historical Document Transcription
Nikolai Vogler, Jonathan Allen, Matthew Miller and Taylor Berg-Kirkpatrick 10:45-12:15 (Regency A & B)

We present a self-supervised pre-training approach for learning rich visual language representations for both handwritten and printed historical document transcription. After supervised fine-tuning of our pre-trained encoder representations for low-resource document transcription on two languages, (1) a heterogeneous set of handwritten Islamicate manuscript images and (2) early modern English printed documents, we show a meaningful improvement in recognition accuracy over the same supervised model trained from scratch with as few as 30 line image transcriptions for training. Our masked language model-style pre-training strategy, where the model is trained to be able to identify the true masked visual representation from distractors sampled from within the same line, encourages learning robust contextualized language representations invariant to scribal writing style and printing noise present across documents.

Hierarchical Transformers Are More Efficient Language Models
Piotr Nawrot, Szymon Tworkowski, Michał Tyrolski, Lukasz Kaiser, Yuhuai Wu, Christian Szegedy and Henryk Michalewski 10:45-12:15 (Regency A & B)
Transformer models yield impressive results on many NLP and sequence modeling tasks. Remarkably, Transformers can handle long sequences, which allows them to produce long coherent outputs: entire paragraphs produced by GPT-3 or well-structured images produced by DALL-E. These large language models are impressive but also very inefficient and costly, which limits their applications and accessibility. We postulate that having an explicit hierarchical architecture is the key to Transformers that efficiently handle long sequences. To verify this claim, we first study different ways to downsample and upsample activations in Transformers so as to make them hierarchical. We use the best performing upsampling and downsampling layers to create Hourglass - a hierarchical Transformer language model. Hourglass improves upon the Transformer baseline given the same amount of computation and can yield the same results as Transformers more efficiently. In particular, Hourglass sets new state-of-the-art for Transformer models on the ImageNet32 generation task and improves language modeling efficiency on the widely studied enwik8 benchmark.

”Diversity and Uncertainty in Moderation” are the Key to Data Selection for Multilingual Few-shot Transfer
Shanu Kumar, Sandipan Dandapat and Monojit Choudhury 10:45-12:15 (Regency A & B)
Few-shot transfer often shows substantial gain over zero-shot transfer [**lauscher2020zero**], which is a practically useful trade-off between fully supervised and unsupervised learning approaches for multilingual pretained model-based systems. This paper explores various strategies for selecting data for annotation that can result in a better few-shot transfer. The proposed approaches rely on multiple measures such as data entropy using n-gram language model, predictive entropy, and gradient embedding. We propose a loss embedding method for sequence labeling tasks, which induces diversity and uncertainty sampling similar to gradient embedding. The proposed data selection strategies are evaluated and compared for POS tagging, NER, and NLI tasks for up to 20 languages. Our experiments show that the gradient and loss embedding-based strategies consistently outperform random data selection baselines, with gains varying with the initial performance of the zero-shot transfer. Furthermore, the proposed method shows similar trends in improvement even when the model is fine-tuned using a lower proportion of the original task-specific labeled training data for zero-shot transfer.

DOCmT5: Document-Level Pretraining of Multilingual Language Models
Chia-Hsuan Lee, Aditya Siddhant, Viresh Ratnakar and Melvin Johnson 10:45-12:15 (Regency A & B)
In this paper, we introduce DOCmT5, a multilingual sequence-to-sequence language model pretrained with large-scale parallel documents. While previous approaches have focused on leveraging sentence-level parallel data, we try to build a general-purpose pretrained model that can understand and generate long documents. We propose a simple and effective pretraining objective - Document reordering Machine Translation (DrMT), in which the input documents that are shuffled and masked need to be translated. DrMT brings consistent improvements over strong baselines on a variety of document-level generation tasks, including over 12 BLEU points for seen-language pair document-level MT, over 7 BLEU points for unseen-language-pair document-level MT and over 3 ROUGE-1 points for seen-language pair cross-lingual summarization. We achieve state-of-the-art (SOTA) on WMT20 De-En and IWSLT15 Zh-En document translation tasks. We also conduct extensive analysis on various factors for document pretraining, including (1) the effects of pretraining data quality and (2) The effects of combining mono-lingual and cross-lingual pretraining. We plan to make our model checkpoints publicly available.

Por Qué Não Utiliser Alla Språk? Mixed Training with Gradient Optimization in Few-Shot Cross-Lingual Transfer
Haoran Xu and Kenton Murray 10:45-12:15 (Regency A & B)
The current state-of-the-art for few-shot cross-lingual transfer learning first trains on abundant labeled data in the source language and then fine-tunes with a few examples on the target language, termed target-adapting. Though this has been demonstrated to work on a variety of tasks, in this paper we show some deficiencies of this approach and propose a one-step mixed training method that trains on both source and target data with stochastic gradient surgery, a novel gradient-level optimization. Unlike the previous studies that focus on one language at a time when target-adapting, we use one model to handle all target languages simultaneously to avoid excessively language-specific models. Moreover, we discuss the unreality of utilizing large target development sets for model selection in previous literature. We further show that our method is both development-free for target languages, and is also able to escape from overfitting issues. We conduct a large-scale experiment on 4 diverse NLP tasks across up to 48 languages. Our proposed method achieves state-of-the-art performance on all tasks and outperforms target-adapting by a large margin, especially for languages that are linguistically distant from the source language, e.g., 7.36% F1 absolute gain on average for the NER task, up to 17.60% on Punjabi.

MTG: A Benchmark Suite for Multilingual Text Generation
Yiran Chen, Zhenqiao Song, Xianze Wu, Danqing Wang, Jingjing Xu, Jiaze Chen, Hao Zhou and Lei Li 10:45-12:15 (Regency A & B)
We introduce MTG, a new benchmark suite for training and evaluating multilingual text generation. It is the first-proposed multilingual multiway text generation dataset with the largest human-annotated data (400k). It includes four generation tasks (story generation, question generation, title generation and text summarization) across five languages (English, German, French, Spanish and Chinese). The multiway setup enables testing knowledge transfer capabilities for a model across languages and tasks. Using MTG, we train and analyze several popular multilingual generation models from different aspects. Our benchmark suite fosters model performance enhancement with more human-annotated parallel data. It provides comprehensive evaluations with diverse generation scenarios. Code and data are available at `https://github.com/zide05/MTG`.

MWP-BERT: Numeracy-Augmented Pre-training for Math Word Problem Solving
Zhenwen Liang, Jipeng Zhang, Lei Wang, Wei Qin, Yunshi Lan, Jie Shao and Xiangliang Zhang 10:45-12:15 (Regency A & B)
Math word problem (MWP) solving faces a dilemma in number representation learning. In order to avoid the number representation issue and reduce the search space of feasible solutions, existing works striving for MWP solving usually replace real numbers with symbolic placeholders to focus on logic reasoning. However, different from common symbolic reasoning tasks like program synthesis and knowledge graph reasoning, MWP solving has extra requirements in numerical reasoning. In other words, instead of the number value itself, it is the reusable numerical property that matters more in numerical reasoning. Therefore, we argue that injecting numerical properties into symbolic placeholders with contextualized representation learning schema can provide a way out of the dilemma in the number representation issue here. In this work, we introduce this idea to the popular pre-training language model (PLM) techniques and build MWP-BERT, an effective contextual number representation PLM. We demonstrate the effectiveness of our MWP-BERT on MWP solving and several MWP-specific understanding tasks on both English and Chinese benchmarks.

Exploiting Numerical-Contextual Knowledge to Improve Numerical Reasoning in Question Answering
Jeonghwan Kim, Junmo Kang, Kyung-min Kim, Giwon Hong and Sung-Hyon Myaeng 10:45-12:15 (Regency A & B)
Numerical reasoning over text is a challenging subtask in question answering (QA) that requires both the understanding of texts and numbers. However, existing language models in these numerical reasoning QA models tend to overly rely on the pre-existing parametric knowledge at inference time, which commonly causes hallucination in interpreting numbers. Our work proposes a novel attention masked reasoning model, the NC-BERT, that learns to leverage the number-related contextual knowledge to alleviate the over-reliance on parametric knowledge and enhance the numerical reasoning capabilities of the QA model. The empirical results suggest that understanding of numbers in their context by reducing the parametric knowledge influence, and refining numerical information in the number embeddings lead to improved numerical reasoning accuracy and performance in DROP, a numerical QA dataset.

METGEN: A Module-Based Entailment Tree Generation Framework for Answer Explanation
Ruixin Hong, Hongming Zhang, Xintong Yu and Changshui Zhang 10:45-12:15 (Regency A & B)
Knowing the reasoning chains from knowledge to the predicted answers can help construct an explainable question answering (QA) system. Advances on QA explanation propose to explain the answers with entailment trees composed of multiple entailment steps. While current work proposes to generate entailment trees with end-to-end generative models, the steps in the generated trees are not constrained and could be unreliable. In this paper, we propose METGEN, a Module-based Entailment Tree GENeration framework that has multiple modules and a reasoning controller. Given a question and several supporting knowledge, METGEN can iteratively generate the entailment tree by conducting single-step entailment with separate modules and selecting the reasoning flow with the controller. As each module is guided to perform a specific type of entailment reasoning, the steps generated by METGEN are more reliable and valid. Experiment results on the standard benchmark show that METGEN can outperform previous state-of-the-art models with only 9% of the parameters.

Challenges in Generalization in Open Domain Question Answering
Linqing Liu, Patrick Lewis, Sebastian Riedel and Pontus Stenetorp 10:45-12:15 (Regency A & B)
Recent work on Open Domain Question Answering has shown that there is a large discrepancy in model performance between novel test questions and those that largely overlap with training questions. However, it is unclear which aspects of novel questions make them challenging. Drawing upon studies on systematic generalization, we introduce and annotate questions according to three categories that measure different levels and kinds of generalization: training set overlap, compositional generalization (comp-gen), and novel-entity generalization (novel-entity). When evaluating six popular parametric and non-parametric models, we find that for the established Natural Questions and TriviaQA datasets, even the strongest model performance for comp-gen/novel-entity is 13.1/5.4% and 9.6/1.5% lower compared to that for the full test set – indicating the challenge posed by these types of questions. Furthermore, we show that whilst non-parametric models can handle questions containing novel entities relatively well, they struggle with those requiring compositional generalization. Lastly, we find that key question difficulty factors are: cascading errors from the retrieval component, frequency of question pattern, and frequency of the entity.

CCQA: A New Web-Scale Question Answering Dataset for Model Pre-Training
Patrick Huber, Armen Aghajanyan, Barlas Oguz, Dmytro Okhonko, Scott Yih, Sonal Gupta and Xilun Chen 10:45-12:15 (Regency A & B)
We propose a novel open-domain question-answering dataset based on the Common Crawl project. With a previously unseen number of around 130 million multilingual question-answer pairs (including about 60 million English data-points), we use our large-scale, natural, diverse and high-quality corpus to in-domain pre-train popular language models for the task of question-answering. In our experiments, we find that our Common Crawl Question Answering dataset (CCQA) achieves promising results in zero-shot, low resource and fine-tuned settings across multiple tasks, models and benchmarks.

UniK-QA: Unified Representations of Structured and Unstructured Knowledge for Open-Domain Question Answering
Barlas Oguz, Xilun Chen, Vladimir Karpukhin, Stan Peshterliev, Dmytro Okhonko, Michael Sejr Schlichtkrull, Sonal Gupta, Yashar Mehdad and Scott Yih 10:45-12:15 (Regency A & B)
We study open-domain question answering with structured, unstructured and semi-structured knowledge sources, including text, tables, lists and knowledge bases. Departing from prior work, we propose a unifying approach that homogenizes all sources by reducing them to text and applies the retriever-reader model which has so far been limited to text sources only. Our approach greatly improves the results on knowledge-base QA tasks by 11 points, compared to latest graph-based methods. More importantly, we demonstrate that our unified knowledge (UniK-QA) model is a simple and yet effective way to combine heterogeneous sources of knowledge, advancing the state-of-the-art results on two popular question answering benchmarks, NaturalQuestions and WebQuestions, by 3.5 and 2.6 points, respectively.

The code of UniK-QA is available at: https://github.com/facebookresearch/UniK-QA.

PerKGQA: Question Answering over Personalized Knowledge Graphs
Ritam Dutt, Kasturi Bhattacharjee, Rashmi Gangadharaiah, Dan Roth and Carolyn Rose 10:45-12:15 (Regency A & B)
Previous studies on question answering over knowledge graphs have typically operated over a single knowledge graph (KG). This KG is assumed to be known a priori and is lever- aged similarly for all users' queries during inference. However, such an assumption is not applicable to real-world settings, such as health- care, where one needs to handle queries of new users over unseen KGs during inference. Furthermore, privacy concerns and high computational costs render it infeasible to query the single KG that has information about all users while answering a specific user's query. The above concerns motivate our question answer- ing setting over personalized knowledge graphs (PERKGQA) where each user has restricted access to their KG. We observe that current state-of-the-art KGQA methods that require learning prior node representations fare poorly. We propose two complementary approaches, PATHCBR and PATHRGCN for PERKGQA. The former is a simple non-parametric technique that employs case-based reasoning, while the latter is a parametric approach using graph neural networks. Our proposed methods circumvent learning prior representations, can generalize to unseen KGs, and outperform strong baselines on an academic and an internal dataset by 6.5% and 10.5%.

TaCL: Improving BERT Pre-training with Token-aware Contrastive Learning
Yixuan Su, Fangyu Liu, Zaiqiao Meng, Tian Lan, Lei Shu, Ehsan Shareghi and Nigel Collier 10:45-12:15 (Regency A & B)
Masked language models (MLMs) such as BERT have revolutionized the field of Natural Language Understanding in the past few years. However, existing pre-trained MLMs often output an anisotropic distribution of token representations that occupies a narrow subset of the entire representation space. Such token representations are not ideal, especially for tasks that demand discriminative semantic meanings of distinct tokens. In this work, we propose TaCL (Token-aware Contrastive Learning), a novel continual pre-training approach that encourages BERT to learn an isotropic and discriminative distribution of token representations. TaCL is fully unsupervised and requires no additional data. We extensively test our approach on a wide range of English and Chinese benchmarks. The results show that TaCL brings consistent and notable improvements over the original BERT model. Furthermore, we conduct detailed analysis to reveal the merits and inner-workings of our approach.

A Question-Answer Driven Approach to Reveal Affirmative Interpretations from Verbal Negations
Md Mosharaf Hossain, Luke Holman, Anusha Kakileti, Tiffany Iris Kao, Nathan Raul Brito, Aaron Abraham Mathews and Eduardo Blanco

10:45-12:15 (Regency A & B)
This paper explores a question-answer driven approach to reveal affirmative interpretations from verbal negations (i.e., when a negation cue grammatically modifies a verb). We create a new corpus consisting of 4,472 verbal negations and discover that 67.1% of them convey that an event actually occurred. Annotators generate and answer 7,277 questions % converted for 4,000 for the 3,001 negations that convey an affirmative interpretation. We first cast the problem of revealing affirmative interpretations from negations as a natural language inference (NLI) classification task. Experimental results show that state-of-the-art transformers trained with existing NLI corpora are insufficient to reveal affirmative interpretations. We also observe, however, that fine-tuning brings substantial improvements. In addition to NLI classification, we also explore the more realistic task of generating affirmative interpretations directly from negations with the T5 transformer. We conclude that the generation task remains a challenge as T5 substantially underperforms humans.

The Role of Context in Detecting Previously Fact-Checked Claims
Shaden Shaar, Firoj Alam, Giovanni Da San Martino and Preslav Nakov 10:45-12:15 (Regency A & B)
Recent years have seen the proliferation of disinformation and fake news online. Traditional approaches to mitigate these issues is to use manual or automatic fact-checking. Recently, another approach has emerged: checking whether the input claim has previously been fact-checked, which can be done automatically, and thus fast, while also offering credibility and explainability, thanks to the human fact-checking and explanations in the associated fact-checking article. Here, we focus on claims made in a political debate and we study the impact of modeling the context of the claim: both on the source side, i.e., in the debate, as well as on the target side, i.e., in the fact-checking explanation document. We do this by modeling the local context, the global context, as well as by means of co-reference resolution, and multi-hop reasoning over the sentences of the document describing the fact-checked claim. The experimental results show that each of these represents a valuable information source, but that modeling the source-side context is most important, and can yield 10+ points of absolute improvement over a state-of-the-art model.

SEQZERO: Few-shot Compositional Semantic Parsing with Sequential Prompts and Zero-shot Models
Jingfeng Yang, Haoming Jiang, Qingyu Yin, Danqing Zhang, Bing Yin and Diyi Yang 10:45-12:15 (Regency A & B)
Recent research showed promising results on combining pretrained language models (LMs) with canonical utterance for few-shot semantic parsing. The canonical utterance is often lengthy and complex due to the compositional structure of formal languages. Learning to generate such canonical utterance requires significant amount of data to reach high performance. Fine-tuning with only few-shot samples, the LMs can easily forget pretrained knowledge, overfit spurious biases, and suffer from compositionally out-of-distribution generalization errors. To tackle these issues, we propose a novel few-shot semantic parsing method – SEQZERO. SEQZERO decomposes the problem into a sequence of sub-problems, which corresponds to the sub-clauses of the formal language. Based on the decomposition, the LMs only need to generate short answers using prompts for predicting sub-clauses. Thus, SEQZERO avoids generating a long canonical utterance at once. Moreover, SEQZERO employs not only a few-shot model but also a zero-shot model to alleviate the overfitting. In particular, SEQZERO brings out the merits from both models via ensemble equipped with our proposed constrained rescaling. SEQZERO achieves SOTA performance of BART-based models on GeoQuery and EcommerceQuery, which are two few-shot datasets with compositional data split.

Weakly Supervised Text-to-SQL Parsing through Question Decomposition
Tomer Wolfson, Daniel Deutch and Jonathan Berant 10:45-12:15 (Regency A & B)
Text-to-SQL parsers are crucial in enabling non-experts to effortlessly query relational data. Training such parsers, by contrast, generally requires expertise in annotating natural language (NL) utterances with corresponding SQL queries. In this work, we propose a weak supervision approach for training text-to-SQL parsers. We take advantage of the recently proposed question meaning representation called QDMR, an intermediate between NL and formal query languages. Given questions, their QDMR structures (annotated by non-experts or automatically predicted), and the answers, we are able to automatically synthesize SQL queries that are used to train text-to-SQL models. We test our approach by experimenting on five benchmark datasets. Our results show that the weakly supervised models perform competitively with those trained on annotated NL-SQL data. Overall, we effectively train text-to-SQL parsers, while using zero SQL annotations.

POLITICS: Pretraining with Same-story Article Comparison for Ideology Prediction and Stance Detection
Yujian Liu, Xinliang Frederick Zhang, David Wegsman, Nicholas Beauchamp and Lu Wang 10:45-12:15 (Regency A & B)
Ideology is at the core of political science research. Yet, there still does not exist general-purpose tools to characterize and predict ideology across different genres of text. To this end, we study Pretrained Language Models using novel ideology-driven pretraining objectives that rely on the comparison of articles on the same story written by media of different ideologies. We further collect a large-scale dataset, consisting of more than 3.6M political news articles, for pretraining. Our model POLITICS outperforms strong baselines and the previous state-of-the-art models on ideology prediction and stance detection tasks. Further analyses show that POLITICS is especially good at understanding long or formally written texts, and is also robust in few-shot learning scenarios.

A Survey on Stance Detection for Mis- and Disinformation Identification
Momchil Hardalov, Arnav Arora, Preslav Nakov and Isabelle Augenstein 10:45-12:15 (Regency A & B)
Understanding attitudes expressed in texts, also known as stance detection, plays an important role in systems for detecting false information online, be it misinformation (unintentionally false) or disinformation (intentionally false information). Stance detection has been framed in different ways, including (a) as a component of fact-checking, rumour detection, and detecting previously fact-checked claims, or (b) as a task in its own right. While there have been prior efforts to contrast stance detection with other related tasks such as argumentation mining and sentiment analysis, there is no existing survey on examining the relationship between stance detection and mis- and disinformation detection. Here, we aim to bridge this gap by reviewing and analysing existing work in this area, with mis- and disinformation in focus, and discussing lessons learnt and future challenges.

Session 10 - 14:15-15:45

Ethics, Bias, Fairness 2

14:15-15:45 (Columbia A)

Selective Differential Privacy for Language Modeling
Weiyan Shi, Aiqi Cui, Evan Li, Ruoxi Jia and Zhou Yu 14:15-14:30 (Columbia A)
With the increasing applications of language models, it has become crucial to protect these models from leaking private information. Previous work has attempted to tackle this challenge by training RNN-based language models with differential privacy guarantees. However, applying classical differential privacy to language models leads to poor model performance as the underlying privacy notion is over-pessimistic and

provides undifferentiated protection for all tokens in the data. Given that the private information in natural language is sparse (for example, the bulk of an email might not carry personally identifiable information), we propose a new privacy notion, selective differential privacy, to provide rigorous privacy guarantees on the sensitive portion of the data to improve model utility. To realize such a new notion, we develop a corresponding privacy mechanism, Selective-DPSGD, for RNN-based language models. Besides language modeling, we also apply the method to a more concrete application – dialog systems. Experiments on both language modeling and dialog system building show that the proposed privacy-preserving mechanism achieves better utilities while remaining safe under various privacy attacks compared to the baselines. The data and code are released at https://github.com/wyshi/lm_privacy to facilitate future research.

Federated Learning with Noisy User Feedback
Rahul Sharma, Anil Ramakrishna, Ansel MacLaughlin, Anna Rumshisky, Jimit Majmudar, Clement Chung, Salman Avestimehr and Rahul Gupta 14:30-14:45 (Columbia A)
Machine Learning (ML) systems are getting increasingly popular, and drive more and more applications and services in our daily life. This has led to growing concerns over user privacy, since human interaction data typically needs to be transmitted to the cloud in order to train and improve such systems. Federated learning (FL) has recently emerged as a method for training ML models on edge devices using sensitive user data and is seen as a way to mitigate concerns over data privacy. However, since ML models are most commonly trained with label supervision, we need a way to extract labels on edge to make FL viable. In this work, we propose a strategy for training FL models using positive and negative user feedback. We also design a novel framework to study different noise patterns in user feedback, and explore how well standard noise-robust objectives can help mitigate this noise when training models in a federated setting. We evaluate our proposed training setup through detailed experiments on two text classification datasets and analyze the effects of varying levels of user reliability and feedback noise on model performance. We show that our method improves substantially over a self-training baseline, achieving performance closer to models trained with full supervision.

Provably Confidential Language Modelling
Xuandong Zhao, Lei Li and Yu-Xiang Wang 14:45-15:00 (Columbia A)
Large language models are shown to memorize privacy information such as social security numbers in training data. Given the sheer scale of the training corpus, it is challenging to screen and filter these privacy data, either manually or automatically. In this paper, we propose Confidentially Redacted Training (CRT), a method to train language generation models while protecting the confidential segments. We borrow ideas from differential privacy (which solves a related but distinct problem) and show that our method is able to provably prevent unintended memorization by randomizing parts of the training process. Moreover, we show that redaction with an approximately correct screening policy amplifies the confidentiality guarantee. We implement the method for both LSTM and GPT language models. Our experimental results show that the models trained by CRT obtain almost the same perplexity while preserving strong confidentiality.

Optimising Equal Opportunity Fairness in Model Training
Aili Shen, Xudong Han, Trevor Cohn, Timothy Baldwin and Lea Frermann 15:00-15:15 (Columbia A)
Real-world datasets often encode stereotypes and societal biases. Such biases can be implicitly captured by trained models, leading to biased predictions and exacerbating existing societal preconceptions. Existing debiasing methods, such as adversarial training and removing protected information from representations, have been shown to reduce bias. However, a disconnect between fairness criteria and training objectives makes it difficult to reason theoretically about the effectiveness of different techniques. In this work, we propose two novel training objectives which directly optimise for the widely-used criterion of *equal opportunity*, and show that they are effective in reducing bias while maintaining high performance over two classification tasks.

How Gender Debiasing Affects Internal Model Representations, and Why It Matters
Hadas Orgad, Seraphina Goldfarb-Tarrant and Yonatan Belinkov 15:15-15:30 (Columbia A)
Common studies of gender bias in NLP focus either on extrinsic bias measured by model performance on a downstream task or on intrinsic bias found in models' internal representations. However, the relationship between extrinsic and intrinsic bias is relatively unknown. In this work, we illuminate this relationship by measuring both quantities together: we debias a model during downstream fine-tuning, which reduces extrinsic bias, and measure the effect on intrinsic bias, which is operationalized as bias extractability with information-theoretic probing. Through experiments on two tasks and multiple bias metrics, we show that our intrinsic bias metric is a better indicator of debiasing than (a contextual adaptation of) the standard WEAT metric, and can also expose cases of superficial debiasing. Our framework provides a comprehensive perspective on bias in NLP models, which can be applied to deploy NLP systems in a more informed manner. Our code and model checkpoints are publicly available.

Explaining Toxic Text via Knowledge Enhanced Text Generation
Rohit Sridhar and Diyi Yang 15:30-15:45 (Columbia A)
Warning: This paper contains content that is offensive and may be upsetting.
Biased or toxic speech can be harmful to various demographic groups. Therefore, it is not only important for models to detect these speech, but to also output explanations of why a given text is toxic. Previous literature has mostly focused on classifying and detecting toxic speech, and existing efforts on explaining stereotypes in toxic speech mainly use standard text generation approaches, resulting in generic and repetitive explanations. Building on these prior works, we introduce a novel knowledge-informed encoder-decoder framework to utilize multiple knowledge sources to generate implications of biased text.
Experiments show that our knowledge informed models outperform prior state-of-the-art models significantly, and can generate detailed explanations of stereotypes in toxic speech compared to baselines, both quantitatively and qualitatively.

Semantics

14:15-15:45 (Columbia C)

Falsesum: Generating Document-level NLI Examples for Recognizing Factual Inconsistency in Summarization
Prasetya Ajie Utama, Joshua Bambrick, Nafise Sadat Moosavi and Iryna Gurevych 14:15-14:30 (Columbia C)
Neural abstractive summarization models are prone to generate summaries that are factually inconsistent with their source documents. Previous work has introduced the task of recognizing such factual inconsistency as a downstream application of natural language inference (NLI). However, state-of-the-art NLI models perform poorly in this context due to their inability to generalize to the target task. In this work, we show that NLI models can be effective for this task when the training data is augmented with high-quality task-oriented examples. We introduce Falsesum, a data generation pipeline leveraging a controllable text generation model to perturb human-annotated summaries, introducing varying types of factual inconsistencies. Unlike previously introduced document-level NLI datasets, our generated dataset contains examples that are diverse and inconsistent yet plausible. We show that models trained on a Falsesum-augmented NLI dataset improve the state-of-the-art

performance across four benchmarks for detecting factual inconsistency in summarization.

Maximum Bayes Smatch Ensemble Distillation for AMR Parsing
Young-Suk Lee, Ramón Fernandez Astudillo, Hoang Thanh Lam, Tahira Naseem, Radu Florian and Salim Roukos 14:30-14:45 (Columbia C)
AMR parsing has experienced an unprecendented increase in performance in the last three years, due to a mixture of effects including architecture improvements and transfer learning. Self-learning techniques have also played a role in pushing performance forward. However, for most recent high performant parsers, the effect of self-learning and silver data augmentation seems to be fading. In this paper we propose to overcome this diminishing returns of silver data by combining Smatch-based ensembling techniques with ensemble distillation. In an extensive experimental setup, we push single model English parser performance to a new state-of-the-art, 85.9 (AMR2.0) and 84.3 (AMR3.0), and return to substantial gains from silver data augmentation. We also attain a new state-of-the-art for cross-lingual AMR parsing for Chinese, German, Italian and Spanish. Finally we explore the impact of the proposed technique on domain adaptation, and show that it can produce gains rivaling those of human annotated data for QALD-9 and achieve a new state-of-the-art for BioAMR.

Curriculum: A Broad-Coverage Benchmark for Linguistic Phenomena in Natural Language Understanding
Zeming Chen and Qiyue Gao 14:45-15:00 (Columbia C)
In the age of large transformer language models, linguistic evaluation play an important role in diagnosing models' abilities and limitations on natural language understanding. However, current evaluation methods show some significant shortcomings. In particular, they do not provide insight into how well a language model captures distinct linguistic skills essential for language understanding and reasoning. Thus they fail to effectively map out the aspects of language understanding that remain challenging to existing models, which makes it hard to discover potential limitations in models and datasets. In this paper, we introduce Curriculum as a new format of NLI benchmark for evaluation of broad-coverage linguistic phenomena. Curriculum contains a collection of datasets that covers 36 types of major linguistic phenomena and an evaluation procedure for diagnosing how well a language model captures reasoning skills for distinct types of linguistic phenomena. We show that this linguistic-phenomena-driven benchmark can serve as an effective tool for diagnosing model behavior and verifying model learning quality. In addition, our experiments provide insight into the limitation of existing benchmark datasets and state-of-the-art models that may encourage future research on re-designing datasets, model architectures, and learning objectives.

Syn2Vec: Synset Colexification Graphs for Lexical Semantic Similarity
John Harvill, Roxana Girju and Mark A. Hasegawa-Johnson 15:00-15:15 (Columbia C)
In this paper we focus on patterns of colexification (co-expressions of form-meaning mapping in the lexicon) as an aspect of lexical-semantic organization, and use them to build large scale synset graphs across BabelNet's typologically diverse set of 499 world languages. We introduce and compare several approaches: monolingual and cross-lingual colexification graphs, popular distributional models, and fusion approaches. The models are evaluated against human judgments on a semantic similarity task for nine languages. Our strong empirical findings also point to the importance of universality of our graph synset embedding representations with no need for any language-specific adaptation when evaluated on the lexical similarity task. The insights of our exploratory investigation of large-scale colexification graphs could inspire significant advances in NLP across languages, especially for tasks involving languages which lack dedicated lexical resources, and can benefit from language transfer from large shared cross-lingual semantic spaces.

WiC = TSV = WSD: On the Equivalence of Three Semantic Tasks
Bradley Hauer and Grzegorz Kondrak 15:15-15:30 (Columbia C)
The Word-in-Context (WiC) task has attracted considerable attention in the NLP community, as demonstrated by the popularity of the recent MCL-WiC SemEval shared task. Systems and lexical resources from word sense disambiguation (WSD) are often used for the WiC task and WiC dataset construction. In this paper, we establish the exact relationship between WiC and WSD, as well as the related task of target sense verification (TSV). Building upon a novel hypothesis on the equivalence of sense and meaning distinctions, we demonstrate through the application of tools from theoretical computer science that these three semantic classification problems can be pairwise reduced to each other, and therefore are equivalent. The results of experiments that involve systems and datasets for both WiC and WSD provide strong empirical evidence that our problem reductions work in practice.

EASE: Entity-Aware Contrastive Learning of Sentence Embedding
Sosuke Nishikawa, Ryokan Ri, Ikuya Yamada, Yoshimasa Tsuruoka and Isao Echizen 15:30-15:45 (Columbia C)
We present EASE, a novel method for learning sentence embeddings via contrastive learning between sentences and their related entities. The advantage of using entity supervision is twofold: (1) entities have been shown to be a strong indicator of text semantics and thus should provide rich training signals for sentence embeddings; (2) entities are defined independently of languages and thus offer useful cross-lingual alignment supervision. We evaluate EASE against other unsupervised models both in monolingual and multilingual settings. We show that EASE exhibits competitive or better performance in English semantic textual similarity (STS) and short text clustering (STC) tasks and it significantly outperforms baseline methods in multilingual settings on a variety of tasks. Our source code, pre-trained models, and newly constructed multi-lingual STC dataset are available at https://github.com/studio-ousia/ease.

Linguistic Theories, Cognitive Modeling, Discourse

14:15-15:45 (Columbia D)

Generalized Quantifiers as a Source of Error in Multilingual NLU Benchmarks
Ruixiang Cui, Daniel Hershcovich and Anders Søgaard 14:15-14:30 (Columbia D)
Logical approaches to representing language have developed and evaluated computational models of quantifier words since the 19th century, but today's NLU models still struggle to capture their semantics. We rely on Generalized Quantifier Theory for language-independent representations of the semantics of quantifier words, to quantify their contribution to the errors of NLU models. We find that quantifiers are pervasive in NLU benchmarks, and their occurrence at test time is associated with performance drops. Multilingual models also exhibit unsatisfying quantifier reasoning abilities, but not necessarily worse for non-English languages. To facilitate directly-targeted probing, we present an adversarial generalized quantifier NLI task (GQNLI) and show that pre-trained language models have a clear lack of robustness in generalized quantifier reasoning.

What company do words keep? Revisiting the distributional semantics of J.R. Firth & Zellig Harris
Mikael Brunila and Jack LaViolette 14:30-14:45 (Columbia D)
The power of word embeddings is attributed to the linguistic theory that similar words will appear in similar contexts. This idea is specifically invoked by noting that "you shall know a word by the company it keeps," a quote from British linguist J.R. Firth who, along with his American colleague Zellig Harris, is often credited with the invention of "distributional semantics." While both Firth and Harris are cited in all major

NLP textbooks and many foundational papers, the content and differences between their theories is seldom discussed. Engaging in a close reading of their work, we discover two distinct and in many ways divergent theories of meaning. One focuses exclusively on the internal workings of linguistic forms, while the other invites us to consider words in new company—not just with other linguistic elements, but also in a broader cultural and situational context. Contrasting these theories from the perspective of current debates in NLP, we discover in Firth a figure who could guide the field towards a more culturally grounded notion of semantics. We consider how an expanded notion of "context" might be modeled in practice through two different strategies: comparative stratification and syntagmatic extension.

Learning the Ordering of Coordinate Compounds and Elaborate Expressions in Hmong, Lahu, and Chinese
Chenxuan Cui, Katherine J. Zhang and David R Mortensen 14:45-15:00 (Columbia D)
Coordinate compounds (CCs) and elaborate expressions (EEs) are coordinate constructions common in languages of East and Southeast Asia. Mortensen (2006) claims that (1) the linear ordering of EEs and CCs in Hmong, Lahu, and Chinese can be predicted via phonological hierarchies and (2) that these phonological hierarchies lack a clear phonetic rationale. These claims are significant because morphosyntax has often been seen as in a feed-forward relationship with phonology, and phonological generalizations have often been assumed to be phonetically "natural". We investigate whether the ordering of CCs and EEs can be learned empirically and whether computational models (classifiers and sequence-labeling models) learn unnatural hierarchies similar to those posited by Mortensen (2006). We find that decision trees and SVMs learn to predict the order of CCs/EEs on the basis of phonology, beating strong baselines for all three languages, with DTs learning hierarchies strikingly similar to those proposed by Mortensen. However, we also find that a neural sequence labeling model is able to learn the ordering of elaborate expressions in Hmong very effectively without using any phonological information. We argue that EE ordering can be learned through two independent routes: phonology and lexical distribution, presenting a more nuanced picture than previous work.

Neural Language Taskonomy: Which NLP Tasks are the most Predictive of fMRI Brain Activity?
Subba Reddy Oota, Jashn Arora, Veeral Agarwal, Mounika Marreddy, Manish Gupta and Bapi Raju Surampudi 15:00-15:15 (Columbia D)
Several popular Transformer based language models have been found to be successful for text-driven brain encoding. However, existing literature leverages only pretrained text Transformer models and has not explored the efficacy of task-specific learned Transformer representations. In this work, we explore transfer learning from representations learned for ten popular natural language processing tasks (two syntactic and eight semantic) for predicting brain responses from two diverse datasets: Pereira (subjects reading sentences from paragraphs) and Narratives (subjects listening to the spoken stories). Encoding models based on task features are used to predict activity in different regions across the whole brain. Features from coreference resolution, NER, and shallow syntax parsing explain greater variance for the reading activity. On the other hand, for the listening activity, tasks such as paraphrase generation, summarization, and natural language inference show better encoding performance. Experiments across all 10 task representations provide the following cognitive insights: (i) language left hemisphere has higher predictive brain activity versus language right hemisphere, (ii) posterior medial cortex, temporo-parieto-occipital junction, dorsal frontal lobe have higher correlation versus early auditory and auditory association cortex, (iii) syntactic and semantic tasks display a good predictive performance across brain regions for reading and listening stimuli resp.

Towards Understanding Large-Scale Discourse Structures in Pre-Trained and Fine-Tuned Language Models
Patrick Huber and Giuseppe Carenini 15:15-15:30 (Columbia D)
In this paper, we extend the line of BERTology work by focusing on the important, yet less explored, alignment of pre-trained and fine-tuned PLMs with large-scale discourse structures. We propose a novel approach to infer discourse information for arbitrarily long documents. In our experiments, we find that the captured discourse information is local and general, even across a collection of fine-tuning tasks. We compare the inferred discourse trees with supervised, distantly supervised and simple baselines to explore the structural overlap, finding that constituency discourse trees align well with supervised models, however, contain complementary discourse information. Lastly, we individually explore self-attention matrices to analyze the information redundancy. We find that similar discourse information is consistently captured in the same heads.

Social Norms Guide Reference Resolution
Mitchell Abrams and Matthias Scheutz 15:30-15:45 (Columbia D)
Humans use natural language, vision, and context to resolve referents in their environment. While some situated reference resolution is trivial, ambiguous cases arise when the language is underspecified or there are multiple candidate referents. This study investigates how pragmatic modulators external to the linguistic content are critical for the correct interpretation of referents in these scenarios. In particular, we demonstrate in a human subjects experiment how the social norms applicable in the given context influence the interpretation of referring expressions. Additionally, we highlight how current coreference tools in natural language processing fail to handle these ambiguous cases. We also briefly discuss the implications of this work for assistive robots which will routinely need to resolve referents in their environment.

Machine Learning 4

14:15-15:45 (Elwha A)

A Structured Span Selector
Tianyu Liu, Yuchen Eleanor Jiang, Ryan Cotterell and Mrinmaya Sachan 14:15-14:30 (Elwha A)
Many natural language processing tasks, e.g., coreference resolution and semantic role labeling, require selecting text spans and making decisions about them. A typical approach to such tasks is to score all possible spans and greedily select spans for task-specific downstream processing. This approach, however, does not incorporate any inductive bias about what sort of spans ought to be selected, e.g., that selected spans tend to be syntactic constituents. In this paper, we propose a novel grammar-based structured span selection model which learns to make use of the partial span-level annotation provided for such problems. Compared to previous approaches, our approach gets rid of the heuristic greedy span selection scheme, allowing us to model the downstream task on an optimal set of spans. We evaluate our model on two popular span prediction tasks: coreference resolution and semantic role labeling; and show improvements on both.

Entity Linking via Explicit Mention-Mention Coreference Modeling
Dhruv Agarwal, Rico Angell, Nicholas Monath and Andrew McCallum 14:30-14:45 (Elwha A)
Learning representations of entity mentions is a core component of modern entity linking systems for both candidate generation and making linking predictions. In this paper, we present and empirically analyze a novel training approach for learning mention and entity representations that is based on building minimum spanning arborescences (i.e., directed spanning trees) over mentions and entities across documents to explicitly model mention coreference relationships. We demonstrate the efficacy of our approach by showing significant improvements in both candidate generation recall and linking accuracy on the Zero-Shot Entity Linking dataset and MedMentions, the largest publicly available biomedical dataset. In addition, we show that our improvements in candidate generation yield higher quality re-ranking models downstream, setting a new SOTA result in linking accuracy on MedMentions. Finally, we demonstrate that our improved mention representations are also

effective for the discovery of new entities via cross-document coreference.

Automatic Multi-Label Prompting: Simple and Interpretable Few-Shot Classification
Han Wang, Canwen Xu and Julian McAuley 14:45-15:00 (Elwha A)
Prompt-based learning (i.e., prompting) is an emerging paradigm for exploiting knowledge learned by a pretrained language model. In this paper, we propose Automatic Multi-Label Prompting (AMuLaP), a simple yet effective method to automatically select label mappings for few-shot text classification with prompting. Our method exploits one-to-many label mappings and a statistics-based algorithm to select label mappings given a prompt template. Our experiments demonstrate that AMuLaP achieves competitive performance on the GLUE benchmark without human effort or external resources.

AcTune: Uncertainty-Based Active Self-Training for Active Fine-Tuning of Pretrained Language Models
Yue Yu, Lingkai Kong, Jieyu Zhang, Rongzhi Zhang and Chao Zhang 15:00-15:15 (Elwha A)
Although fine-tuning pre-trained language models (PLMs) renders strong performance in many NLP tasks, it relies on excessive labeled data. Recently, researchers have resorted to active fine-tuning for enhancing the label efficiency of PLM fine-tuning, but existing methods of this type usually ignore the potential of unlabeled data. We develop AcTune, a new framework that improves the label efficiency of active PLM fine-tuning by unleashing the power of unlabeled data via self-training. AcTune switches between data annotation and model self-training based on uncertainty: the unlabeled samples of high-uncertainty are selected for annotation, while the ones from low-uncertainty regions are used for model self-training. Additionally, we design (1) a region-aware sampling strategy to avoid redundant samples when querying annotations and (2) a momentum-based memory bank to dynamically aggregate the model's pseudo labels to suppress label noise in self-training. Experiments on 6 text classification datasets show that AcTune outperforms the strongest active learning and self-training baselines and improves the label efficiency of PLM fine-tuning by 56.2% on average. Our implementation is available at `https://github.com/yueyu1030/actune`.

MoEBERT: from BERT to Mixture-of-Experts via Importance-Guided Adaptation
Simiao Zuo, Qingru Zhang, Chen Liang, Pengcheng He, Tuo Zhao and Weizhu Chen 15:15-15:30 (Elwha A)
Pre-trained language models have demonstrated superior performance in various natural language processing tasks. However, these models usually contain hundreds of millions of parameters, which limits their practicality because of latency requirements in real-world applications. Existing methods train small compressed models via knowledge distillation. However, performance of these small models drops significantly compared with the pre-trained models due to their reduced model capacity. We propose MoEBERT, which uses a Mixture-of-Experts structure to increase model capacity and inference speed. We initialize MoEBERT by adapting the feed-forward neural networks in a pre-trained model into multiple experts. As such, representation power of the pre-trained model is largely retained. During inference, only one of the experts is activated, such that speed can be improved. We also propose a layer-wise distillation method to train MoEBERT. We validate the efficiency and efficacy of MoEBERT on natural language understanding and question answering tasks. Results show that the proposed method outperforms existing task-specific distillation algorithms. For example, our method outperforms previous approaches by over 2% on the MNLI (mismatched) dataset. Our code is publicly available at https://github.com/SimiaoZuo/MoEBERT.

Teaching BERT to Wait: Balancing Accuracy and Latency for Streaming Disfluency Detection
Angelica Chen, Vicky Zayats, Daniel David Walker and Dirk Padfield 15:30-15:45 (Elwha A)
In modern interactive speech-based systems, speech is consumed and transcribed incrementally prior to having disfluencies removed. While this post-processing step is crucial for producing clean transcripts and high performance on downstream tasks (e.g. machine translation), most current state-of-the-art NLP models such as the Transformer operate non-incrementally, potentially causing unacceptable delays for the user. In this work we propose a streaming BERT-based sequence tagging model that, combined with a novel training objective, is capable of detecting disfluencies in real-time while balancing accuracy and latency. This is accomplished by training the model to decide whether to immediately output a prediction for the current input or to wait for further context, in essence learning to dynamically size the lookahead window. Our results demonstrate that our model produces comparably accurate predictions and does so sooner than our baselines, with lower flicker. Furthermore, the model attains state-of-the-art latency and stability scores when compared with recent work on incremental disfluency detection.

Question Answering 2

14:15-15:45 (Elwha B)

Learning to Retrieve Passages without Supervision
Ori Ram, Gal Shachaf, Omer Levy, Jonathan Berant and Amir Globerson 14:15-14:30 (Elwha B)
Dense retrievers for open-domain question answering (ODQA) have been shown to achieve impressive performance by training on large datasets of question-passage pairs. In this work we ask whether this dependence on labeled data can be reduced via unsupervised pretraining that is geared towards ODQA. We show this is in fact possible, via a novel pretraining scheme designed for retrieval. Our "recurring span retrieval" approach uses recurring spans across passages in a document to create pseudo examples for contrastive learning. Our pretraining scheme directly controls for term overlap across pseudo queries and relevant passages, thus allowing to model both lexical and semantic relations between them. The resulting model, named Spider, performs surprisingly well without any labeled training examples on a wide range of ODQA datasets. Specifically, it significantly outperforms all other pretrained baselines in a zero-shot setting, and is competitive with BM25, a strong sparse baseline. Moreover, a hybrid retriever over Spider and BM25 improves over both, and is often competitive with DPR models, which are trained on tens of thousands of examples. Last, notable gains are observed when using Spider as an initialization for supervised training.

Interpretable Proof Generation via Iterative Backward Reasoning
Hanhao Qu, Yu Cao, Jun Gao, Liang Ding and Ruifeng Xu 14:30-14:45 (Elwha B)
We present IBR, an Iterative Backward Reasoning model to solve the proof generation tasks on rule-based Question Answering (QA), where models are required to reason over a series of textual rules and facts to find out the related proof path and derive the final answer. We handle the limitations of existed works in two folds: 1) enhance the interpretability of reasoning procedures with detailed tracking, by predicting nodes and edges in the proof path iteratively backward from the question; 2) promote the efficiency and accuracy via reasoning on the elaborate representations of nodes and history paths, without any intermediate texts that may introduce external noise during proof generation. There are three main modules in IBR, QA and proof strategy prediction to obtain the answer and offer guidance for the following procedure; parent node prediction to determine a node in the existing proof that a new child node will link to; child node prediction to find out which new node will be added to the proof. Experiments on both synthetic and paraphrased datasets demonstrate that IBR has better in-domain performance as well as cross-domain transferability than several strong baselines. Our code and models are available at https://github. com/find-knowledge/IBR.

MultiSpanQA: A Dataset for Multi-Span Question Answering
Haonan Li, Martin Tomko, Maria Vasardani and Timothy Baldwin 14:45-15:00 (Elwha B)
Most existing reading comprehension datasets focus on single-span answers, which can be extracted as a single contiguous span from a given text passage. Multi-span questions, i.e., questions whose answer is a series of multiple discontiguous spans in the text, are common real life but are less studied. In this paper, we present MultiSpanQA, a new dataset that focuses on multi-span questions. Raw questions and contexts are extracted from the Natural Questions dataset. After multi-span re-annotation, MultiSpanQA consists of over a total of 6,000 multi-span questions in the basic version, and over 19,000 examples with unanswerable questions, and questions with single-, and multi-span answers in the expanded version. We introduce new metrics for the purposes of multi-span question answering evaluation, and establish several baselines using advanced models. Finally, we propose a new model which beats all baselines and achieves state-of-the-art on our dataset.

Evidentiality-guided Generation for Knowledge-Intensive NLP Tasks
Akari Asai, Matt Gardner and Hannaneh Hajishirzi 15:00-15:15 (Elwha B)
Retrieval-augmented generation models have shown state-of-the-art performance across many knowledge-intensive NLP tasks such as open-domain question answering and fact verification. These models are trained to generate a final output given retrieved passages that can be irrelevant to an input query, leading to learning spurious cues or memorization. This work introduces a method to incorporate *evidentiality* of passages—whether a passage contains correct evidence to support the output—into training the generator. We introduce a multi-task learning framework to jointly generate the final output and predict the *evidentiality* of each passage. Furthermore, we introduce a new task-agnostic method for obtaining high-quality *silver* evidentiality labels, addressing the issues of gold evidentiality labels being unavailable in most domains. Our experiments on five datasets across three knowledge-intensive tasks show that our new evidentiality-guided generator significantly outperforms its direct counterpart on all of them, and advances the state of the art on three of them. Our analysis shows that multi-task learning and silver evidentiality mining play key roles. Our code is available at https://github.com/AkariAsai/evidentiality_qa

Modularized Transfer Learning with Multiple Knowledge Graphs for Zero-shot Commonsense Reasoning
Yu Jin Kim, Beong-woo Kwak, Youngwook Kim, Reinald Kim Amplayo, Seung-won Hwang and Jinyoung Yeo 15:15-15:30 (Elwha B)
Commonsense reasoning systems should be able to generalize to diverse reasoning cases. However, most state-of-the-art approaches depend on expensive data annotations and overfit to a specific benchmark without learning how to perform general semantic reasoning. To overcome these drawbacks, zero-shot QA systems have shown promise as a robust learning scheme by transforming a commonsense knowledge graph (KG) into synthetic QA-form samples for model training. Considering the increasing type of different commonsense KGs, this paper aims to extend the zero-shot transfer learning scenario into multiple-source settings, where different KGs can be utilized synergetically. Towards this goal, we propose to mitigate the loss of knowledge from the interference among the different knowledge sources, by developing a modular variant of the knowledge aggregation as a new zero-shot commonsense reasoning framework. Results on five commonsense reasoning benchmarks demonstrate the efficacy of our framework, improving the performance with multiple KGs.

[TACL] **MuSiQue: Multi-hop Questions via Single-hop Question Composition**
Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot and Ashish Sabharwal 15:30-15:45 (Elwha B)
Multi-hop reasoning remains an elusive goal as existing multi-hop benchmarks are known to be largely solvable via shortcuts. Can we create a question answering (QA) dataset that, by construction, *requires* proper multi-hop reasoning? To this end, we introduce a bottom-up approach that systematically selects composable pairs of single-hop questions that are connected, i.e., where one reasoning step critically relies on information from another. This bottom-up methodology lets us explore a vast space of questions and add stringent filters as well as other mechanisms targeting connected reasoning. It provides fine-grained control over the construction process and the properties of the resulting k-hop questions. We use this methodology to create MuSiQue-Ans, a new multihop QA dataset with 25K 2-4 hop questions. Relative to existing datasets, MuSiQue-Ans is more difficult overall (3x increase in human-machine gap), and harder to cheat via disconnected reasoning (e.g., a single-hop model has a 30 point drop in F1). We further add unanswerable contrast questions to produce a more stringent dataset, MuSiQue-Full. We hope our datasets will help the NLP community develop models that perform genuine multi-hop reasoning.

Findings Poster Session 2

14:15-15:45 (Regency A & B)

Opportunities for Human-centered Evaluation of Machine Translation Systems
Daniel J. Liebling, Katherine A Heller, Samantha Robertson and Wesley Deng 14:15-15:45 (Regency A & B)
Machine translation models are embedded in larger user-facing systems. Although model evaluation has matured, evaluation at the systems level is still lacking. We review literature from both the translation studies and HCI communities about who uses machine translation and for what purposes. We emphasize an important difference in evaluating machine translation models versus the physical and cultural systems in which they are embedded. We then propose opportunities for improved measurement of user-facing translation systems. We pay particular attention to the need for design and evaluation to aid engendering trust and enhancing user agency in future machine translation systems.

One Size Does Not Fit All: The Case for Personalised Word Complexity Models
Sian Gooding and Manuel Tragut 14:15-15:45 (Regency A & B)
Complex Word Identification (CWI) aims to detect words within a text that a reader may find difficult to understand. It has been shown that CWI systems can improve text simplification, readability prediction and vocabulary acquisition modelling. However, the difficulty of a word is a highly idiosyncratic notion that depends on a reader's first language, proficiency and reading experience. In this paper, we show that personal models are best when predicting word complexity for individual readers. We use a novel active learning framework that allows models to be tailored to individuals and release a dataset of complexity annotations and models as a benchmark for further research.

Aligning Generative Language Models with Human Values
Ruibo Liu, Ge Zhang, Xinyu Feng and Soroush Vosoughi 14:15-15:45 (Regency A & B)
Although current large-scale generative language models (LMs) can show impressive insights about factual knowledge, they do not exhibit similar success with respect to human values judgements (e.g., whether or not the generations of an LM are moral). Existing methods learn human values either by directly mimicking the behavior of human data, or rigidly constraining the generation space to human-chosen tokens. These methods are inherently limited in that they do not consider the contextual and abstract nature of human values and as a result often fail when dealing with out-of-domain context or sophisticated and abstract human values.

This paper proposes SENSEI, a new reinforcement learning based method that can embed human values judgements into each step of language generation. SENSEI deploys an Actor-Critic framework, where the Critic is a reward distributor that simulates the reward assignment procedure of humans, while the Actor guides the generation towards the maximum reward direction. Compared with five existing methods in three human values alignment datasets, SENSEI not only achieves higher alignment performance in terms of both automatic and human evaluations, but also shows improvements on robustness and transfer learning on unseen human values.

Design Challenges for a Multi-Perspective Search Engine
Sihao Chen, Siyi Liu, Xander Uyttendaele, Yi Zhang, William Bruno and Dan Roth 14:15-15:45 (Regency A & B)
Many users turn to document retrieval systems (e.g. search engines) to seek answers to controversial or open-ended questions. However, classical document retrieval systems fall short at delivering users a set of direct and diverse responses in such cases, which requires identifying responses within web documents in the context of the query, and aggregating the responses based on their different perspectives.

The goal of this work is to survey and study the user information needs for building a multi-perspective search engine of such. We examine the challenges of synthesizing such language understanding objectives with document retrieval, and study a new *perspective-oriented* document retrieval paradigm. We discuss and assess the inherent natural language understanding challenges one needs to address in order to achieve the goal. Following the design challenges and principles, we propose and evaluate a practical prototype pipeline system. We use the prototype system to conduct a user survey in order to assess the utility of our paradigm, as well as understanding the user information needs when issuing controversial and open-ended queries to a search engine.

BehancePR: A Punctuation Restoration Dataset for Livestreaming Video Transcript
Viet Dac Lai, Amir Pouran Ben Veyseh, Franck Dernoncourt and Thien Huu Nguyen 14:15-15:45 (Regency A & B)
Given the increasing number of livestreaming videos, automatic speech recognition and post-processing for livestreaming video transcripts are crucial for efficient data management as well as knowledge mining. A key step in this process is punctuation restoration which restores fundamental text structures such as phrase and sentence boundaries from the video transcripts. This work presents a new human-annotated corpus, called BehancePR, for punctuation restoration in livestreaming video transcripts. Our experiments on BehancePR demonstrate the challenges of punctuation restoration for this domain. Furthermore, we show that popular natural language processing toolkits like Stanford Stanza, Spacy, and Trankit underperform on detecting sentence boundary on non-punctuated transcripts of livestreaming videos. The dataset is publicly accessible at http://github.com/nlp-uoregon/behancepr.

Masked Summarization to Generate Factually Inconsistent Summaries for Improved Factual Consistency Checking
Hwanhee Lee, Kang Min Yoo, Joonsuk Park, Hwaran Lee and Kyomin Jung 14:15-15:45 (Regency A & B)
Despite the recent advances in abstractive summarization systems, it is still difficult to determine whether a generated summary is factual consistent with the source text. To this end, the latest approach is to train a factual consistency classifier on factually consistent and inconsistent summaries. Luckily, the former is readily available as reference summaries in existing summarization datasets. However, generating the latter remains a challenge, as they need to be factually inconsistent, yet closely relevant to the source text to be effective. In this paper, we propose to generate factually inconsistent summaries using source texts and reference summaries with key information masked. Experiments on seven benchmark datasets demonstrate that factual consistency classifiers trained on summaries generated using our method generally outperform existing models and show a competitive correlation with human judgments. We also analyze the characteristics of the summaries generated using our method. We will release the pre-trained model and the code at `https://github.com/hwanheelee1993/MFMA`.

Efficient Few-Shot Fine-Tuning for Opinion Summarization
Arthur Brazinskas, Ramesh Nallapati, Mohit Bansal and Markus Dreyer 14:15-15:45 (Regency A & B)
Abstractive summarization models are typically pre-trained on large amounts of generic texts, then fine-tuned on tens or hundreds of thousands of annotated samples. However, in opinion summarization, large annotated datasets of reviews paired with reference summaries are not available and would be expensive to create. This calls for fine-tuning methods robust to overfitting on small datasets. In addition, generically pre-trained models are often not accustomed to the specifics of customer reviews and, after fine-tuning, yield summaries with disfluencies and semantic mistakes. To address these problems, we utilize an efficient few-shot method based on adapters which, as we show, can easily store in-domain knowledge. Instead of fine-tuning the entire model, we add adapters and pre-train them in a task-specific way on a large corpus of unannotated customer reviews, using held-out reviews as pseudo summaries. Then, fine-tune the adapters on the small available human-annotated dataset. We show that this self-supervised adapter pre-training improves summary quality over standard fine-tuning by 2.0 and 1.3 ROUGE-L points on the Amazon and Yelp datasets, respectively. Finally, for summary personalization, we condition on aspect keyword queries, automatically created from generic datasets. In the same vein, we pre-train the adapters in a query-based manner on customer reviews and then fine-tune them on annotated datasets. This results in better-organized summary content reflected in improved coherence and fewer redundancies.

Make The Most of Prior Data: A Solution for Interactive Text Summarization with Preference Feedback
*Duy-Hung Nguyen, Nguyen Viet Dung Nghiem, Bao-Sinh Nguyen, Dung Tien Tien Le, Shahab Sabahi, Minh-Tien Nguyen and Hung Le*14:15-15:45 (Regency A & B)
For summarization, human preferences is critical to tame outputs of the summarizer in favor of human interests, as ground-truth summaries are scarce and ambiguous. Practical settings require dynamic exchanges between humans and AI agents wherein feedback is provided in an online manner, a few at a time. In this paper, we introduce a new framework to train summarization models with preference feedback interactively. By properly leveraging offline data and a novel reward model, we improve the performance regarding ROUGE scores and sample-efficiency. Our experiments on three various datasets confirm the benefit of the proposed framework in active, few-shot and online settings of preference learning.

PromptGen: Automatically Generate Prompts using Generative Models
Yue Zhang, Hongliang Fei, Dingcheng Li and Ping Li 14:15-15:45 (Regency A & B)
Recently, prompt learning has received significant attention, where the downstream tasks are reformulated to the mask-filling task with the help of a textual prompt. The key point of prompt learning is finding the most appropriate prompt. This paper proposes a novel model `PromptGen`, which can automatically generate prompts conditional on the input sentence. `PromptGen` is the first work considering dynamic prompt generation for knowledge probing, based on a pre-trained generative model. To mitigate any label information leaking from the pre-trained generative model, when given a generated prompt, we replace the query input with “None”. We pursue that this perturbed context-free prompt cannot trigger the correct label. We evaluate our model on the knowledge probing LAMA benchmark, and show that PromptGen significantly outperforms other baselines.

Extracting Temporal Event Relation with Syntax-guided Graph Transformer
Shuaicheng Zhang, Qiang Ning and Lifu Huang 14:15-15:45 (Regency A & B)
Extracting temporal relations (e.g., before, after, and simultaneous) among events is crucial to natural language understanding. One of the key challenges of this problem is that when the events of interest are far away in text, the context in-between often becomes complicated, making it challenging to resolve the temporal relationship between them. This paper thus proposes a new Syntax-guided Graph Transformer network (SGT) to mitigate this issue, by (1) explicitly exploiting the connection between two events based on their dependency parsing trees, and (2) automatically locating temporal cues between two events via a novel syntax-guided attention mechanism. Experiments on two benchmark datasets, MATRES and TB-DENSE, show that our approach significantly outperforms previous state-of-the-art methods on both end-to-end temporal relation extraction and temporal relation classification with up to 7.9% absolute F-score gain; This improvement also proves to be

robust on the contrast set of MATRES. We will make all the programs publicly available once the paper is accepted.

StATIK: Structure and Text for Inductive Knowledge Graph Completion
Elan Sopher Markowitz, Keshav Balasubramanian, Mehrnoosh Mirtaheri, Murali Annavaram, Aram Galstyan and Greg Ver Steeg 14:15-15:45 (Regency A & B)
Knowledge graphs (KGs) often represent knowledge bases that are incomplete. Machine learning models can alleviate this by helping automate graph completion. Recently, there has been growing interest in completing knowledge bases that are dynamic, where previously unseen entities may be added to the KG with many missing links. In this paper, we present **StATIK**–**St**ructure **A**nd **T**ext for **I**nductive **K**nowledge Completion. `StATIK` uses Language Models to extract the semantic information from text descriptions, while using Message Passing Neural Networks to capture the structural information. `StATIK` achieves state of the art results on three challenging inductive baselines. We further analyze our hybrid model through detailed ablation studies.

Permutation Invariant Strategy Using Transformer Encoders for Table Understanding
Sarthak Dash, Sugato Bagchi, Nandana Mihindukulasooriya and Alfio Gliozzo 14:15-15:45 (Regency A & B)
Representing text in tables is essential for many business intelligence tasks such as semantic retrieval, data exploration and visualization, and question answering. Existing methods that leverage pretrained Transformer encoders range from a simple construction of pseudo-sentences by concatenating text across rows or columns to complex parameter-intensive models that encode table structure and require additional pretraining. In this work, we introduce a novel encoding strategy for Transformer encoders that preserves the critical property of permutation invariance across rows or columns. Unlike existing state-of-the-art methods for Table Understanding, our proposed approach does not require any additional pretraining and still substantially outperforms existing methods in almost all instances. We demonstrate the effectiveness of our proposed approach on three table interpretation tasks: column type annotation, relation extraction, and entity linking through extensive experiments on existing tabular datasets.

Self-Training with Differentiable Teacher
Simiao Zuo, Yue Yu, Chen Liang, Haoming Jiang, Siawpeng Er, Chao Zhang, Tuo Zhao and Hongyuan Zha 14:15-15:45 (Regency A & B)
Self-training achieves enormous success in various semi-supervised and weakly-supervised learning tasks. The method can be interpreted as a teacher-student framework, where the teacher generates pseudo-labels, and the student makes predictions. The two models are updated alternatingly. However, such a straightforward alternating update rule leads to training instability. This is because a small change in the teacher may result in a significant change in the student. To address this issue, we propose DRIFT, short for differentiable self-training, that treats teacher-student as a Stackelberg game. In this game, a leader is always in a more advantageous position than a follower. In self-training, the student contributes to the prediction performance, and the teacher controls the training process by generating pseudo-labels. Therefore, we treat the student as the leader and the teacher as the follower. The leader procures its advantage by acknowledging the follower's strategy, which involves differentiable pseudo-labels and differentiable sample weights. Consequently, the leader-follower interaction can be effectively captured via Stackelberg gradient, obtained by differentiating the follower's strategy. Experimental results on semi- and weakly-supervised classification and named entity recognition tasks show that our model outperforms existing approaches by large margins.

Low-resource Entity Set Expansion: A Comprehensive Study on User-generated Text
Yutong Shao, Nikita Bhutani, Sajjadur Rahman and Estevam Hruschka 14:15-15:45 (Regency A & B)
Entity set expansion (ESE) aims at obtaining a more complete set of entities given a textual corpus and a seed set of entities of a concept. Although it is a critical task in many NLP applications, existing benchmarks are limited to well-formed text (e.g., Wikipedia) and well-defined concepts (e.g., countries and diseases). Furthermore, only a small number of predictions are evaluated compared to the actual size of an entity set. A rigorous assessment of ESE methods warrants more comprehensive benchmarks and evaluation. In this paper, we consider user-generated text to understand the generalizability of ESE methods. We develop new benchmarks and propose more rigorous evaluation metrics for assessing the performance of ESE methods. Additionally, we identify phenomena such as non-named entities, multifaceted entities, vague concepts that are more prevalent in user-generated text than well-formed text, and use them to profile ESE methods. We observe that the strong performance of state-of-the-art ESE methods does not generalize well to user-generated text. We conduct comprehensive empirical analysis and draw insights from the findings.

Zero-shot Entity Linking with Less Data
G P Shrivatsa Bhargav, Dinesh Khandelwal, Saswati Dana, Dinesh Garg, Pavan Kapanipathi, Salim Roukos, Alexander Gray and L Venkata Subramaniam 14:15-15:45 (Regency A & B)
Entity Linking (EL) maps an entity mention in a natural language sentence to an entity in a knowledge base (KB). The Zero-shot Entity Linking (ZEL) extends the scope of EL to unseen entities at the test time without requiring new labeled data. BLINK (BERT-based) is one of the SOTA models for ZEL. Interestingly, we discovered that BLINK exhibits diminishing returns, i.e., it reaches 98% of its performance with just 1% of the training data and the remaining 99% of the data yields only a marginal increase of 2% in the performance. While this extra 2% gain makes a huge difference for downstream tasks, training BLINK on large amounts of data is very resource-intensive and impractical. In this paper, we propose a neuro-symbolic, multi-task learning approach to bridge this gap. Our approach boosts the BLINK's performance with much less data by exploiting an auxiliary information about entity types. Specifically, we train our model on two tasks simultaneously - entity linking (primary task) and hierarchical entity type prediction (auxiliary task). The auxiliary task exploits the hierarchical structure of entity types. Our approach achieves superior performance on ZEL task with significantly less training data. On four different benchmark datasets, we show that our approach achieves significantly higher performance than SOTA models when they are trained with just 0.01%, 0.1%, or 1% of the original training data. Our code is available at https://github.com/IBM/NeSLET.

Event Detection for Suicide Understanding
Luis Fernando Guzman-Nateras, Viet Dac Lai, Amir Pouran Ben Veyseh, Franck Dernoncourt and Thien Huu Nguyen 14:15-15:45 (Regency A & B)
Suicide is a serious problem in every society. Understanding life events of a potential patient is essential for successful suicide-risk assessment and prevention. In this work, we focus on the Event Detection (ED) task to identify event trigger words of suicide-related events in public posts of discussion forums. In particular, we introduce SuicideED: a new dataset for the ED task that features seven suicidal event types to comprehensively capture suicide actions and ideation, and general risk and protective factors. Our experiments with current state-of-the-art ED systems suggest that this domain poses meaningful challenges as there is significant room for improvement of ED models. We will release SuicideED to support future research in this important area.

Textual Entailment for Event Argument Extraction: Zero- and Few-Shot with Multi-Source Learning
Oscar Sainz, Itziar Gonzalez-Dios, Oier Lopez De Lacalle, Bonan Min and Eneko Agirre 14:15-15:45 (Regency A & B)
Recent work has shown that NLP tasks such as Relation Extraction (RE) can be recasted as a Textual Entailment tasks using verbalizations, with strong performance in zero-shot and few-shot settings thanks to pre-trained entailment models. The fact that relations in current RE datasets are easily verbalized casts doubts on whether entailment would be effective in more complex tasks. In this work we show that entailment is also effective in Event Argument Extraction (EAE), reducing the need of manual annotation to 50% and 20% in ACE and WikiEvents,

respectively, while achieving the same performance as with full training. More importantly, we show that recasting EAE as entailment alleviates the dependency on schemas, which has been a roadblock for transferring annotations between domains. Thanks to entailment, the multi-source transfer between ACE and WikiEvents further reduces annotation down to 10% and 5% (respectively) of the full training without transfer.Our analysis shows that key to good results is the use of several entailment datasets to pre-train the entailment model. Similar to previous approaches, our method requires a small amount of effort for manual verbalization: only less than 15 minutes per event argument types is needed; comparable results can be achieved from users of different level of expertise.

EA^{2}E: Improving Consistency with Event Awareness for Document-Level Argument Extraction
Qi Zeng, Qiusi Zhan and Heng Ji 14:15-15:45 (Regency A & B)
Events are inter-related in documents. Motivated by the one-sense-per-discourse theory, we hypothesize that a participant tends to play consistent roles across multiple events in the same document. However recent work on document-level event argument extraction models each individual event in isolation and therefore causes inconsistency among extracted arguments across events, which will further cause discrepancy for downstream applications such as event knowledge base population, question answering, and hypothesis generation. In this work, we formulate event argument consistency as the constraints from event-event relations under the document-level setting. To improve consistency we introduce the Event-Aware Argument Extraction (EA^{2}E) model with augmented context for training and inference. Experiment results on WIKIEVENTS and ACE2005 datasets demonstrate the effectiveness of EA^{2}E compared to baseline methods.

Dangling-Aware Entity Alignment with Mixed High-Order Proximities
Juncheng Liu, Zequn Sun, Bryan Hooi, Yiwei Wang, Dayiheng Liu, Baosong Yang, Xiaokui Xiao and Muhao Chen 14:15-15:45 (Regency A & B)
We study dangling-aware entity alignment in knowledge graphs (KGs), which is an underexplored but important problem. As different KGs are naturally constructed by different sets of entities, a KG commonly contains some dangling entities that cannot find counterparts in other KGs. Therefore, dangling-aware entity alignment is more realistic than the conventional entity alignment where prior studies simply ignore dangling entities. We propose a framework using mixed high-order proximities on dangling-aware entity alignment. Our framework utilizes both the local high-order proximity in a nearest neighbor subgraph and the global high-order proximity in an embedding space for both dangling detection and entity alignment. Extensive experiments with two evaluation settings shows that our method more precisely detects dangling entities, and better aligns matchable entities. Further investigations demonstrate that our framework can mitigate the hubness problem on dangling-aware entity alignment.

Literature-Augmented Clinical Outcome Prediction
Aakanksha Naik, Sravanthi Parasa, Sergey Feldman, Lucy Lu Wang and Tom Hope 14:15-15:45 (Regency A & B)
We present BEEP (Biomedical Evidence-Enhanced Predictions), a novel approach for clinical outcome prediction that retrieves patient-specific medical literature and incorporates it into predictive models. Based on each individual patient's clinical notes, we train language models (LMs) to find relevant papers and fuse them with information from notes to predict outcomes such as in-hospital mortality. We develop methods to retrieve literature based on noisy, information-dense patient notes, and to augment existing outcome prediction models with retrieved papers in a manner that maximizes predictive accuracy. Our approach boosts predictive performance on three important clinical tasks in comparison to strong recent LM baselines, increasing F1 by up to 5 points and precision@Top-K by a large margin of over 25%.

Retrieval-Augmented Multilingual Keyphrase Generation with Retriever-Generator Iterative Training
Yifan Gao, Qingyu Yin, Zheng li, Rui Meng, Tong Zhao, Bing Yin, Irwin King and Michael Lyu 14:15-15:45 (Regency A & B)
Keyphrase generation is the task of automatically predicting keyphrases given a piece of long text. Despite its recent flourishing, keyphrase generation on non-English languages haven't been vastly investigated. In this paper, we call attention to a new setting named multilingual keyphrase generation and we contribute two new datasets, EcommerceMKP and AcademicMKP, covering six languages. Technically, we propose a retrieval-augmented method for multilingual keyphrase generation to mitigate the data shortage problem in non-English languages. The retrieval-augmented model leverages keyphrase annotations in English datasets to facilitate generating keyphrases in low-resource languages. Given a non-English passage, a cross-lingual dense passage retrieval module finds relevant English passages. Then the associated English keyphrases serve as external knowledge for keyphrase generation in the current language. Moreover, we develop a retriever-generator iterative training algorithm to mine pseudo parallel passage pairs to strengthen the cross-lingual passage retriever. Comprehensive experiments and ablations show that the proposed approach outperforms all baselines.

Controllable Sentence Simplification via Operation Classification
Liam Cripwell, Joël Legrand and Claire Gardent 14:15-15:45 (Regency A & B)
Different types of transformations have been used to model sentence simplification ranging from mainly local operations such as phrasal or lexical rewriting, deletion and re-ordering to the more global affecting the whole input sentence such as sentence rephrasing, copying and splitting. In this paper, we propose a novel approach to sentence simplification which encompasses four global operations: whether to rephrase or copy and whether to split based on syntactic or discourse structure. We create a novel dataset that can be used to train highly accurate classification systems for these four operations. We propose a controllable-simplification model that tailors simplifications to these operations and show that it outperforms both end-to-end, non-controllable approaches and previous controllable approaches.

The Case for a Single Model that can Both Generate Continuations and Fill-in-the-Blank
Daphne Ippolito, Liam Dugan, Emily Reif, Ann Yuan, Andy Coenen and Chris Callison-Burch 14:15-15:45 (Regency A & B)
The task of inserting text into a specified position in a passage, known as fill in the blank (FitB), is useful for a variety of applications where writers interact with a natural language generation (NLG) system to craft text. While previous work has tackled this problem with models trained specifically to do fill in the blank, a more useful model is one that can effectively perform _both_ FitB and continuation tasks. In this work, we evaluate the feasibility of using a single model to do both tasks. We show that models pre-trained with a FitB-style objective are capable of both tasks, while models pre-trained for continuation are not. Finally, we show how these models can be easily finetuned to allow for fine-grained control over the length and word choice of the generation.

Probing the Role of Positional Information in Vision-Language Models
Philipp J. Rösch and Jindřich Libovický 14:15-15:45 (Regency A & B)
In most Vision-Language models (VL), the understanding of the image structure is enabled by injecting the position information (PI) about objects in the image. In our case study of LXMERT, a state-of-the-art VL model, we probe the use of the PI in the representation and study its effect on Visual Question Answering. We show that the model is not capable of leveraging the PI for the image-text matching task on a challenge set where only position differs. Yet, our experiments with probing confirm that the PI is indeed present in the representation. We introduce two strategies to tackle this: (i) Positional Information Pre-training and (ii) Contrastive Learning on PI using Cross-Modality Matching. Doing so, the model can correctly classify if images with detailed PI statements match. Additionally to the 2D information from bounding boxes, we introduce the object's depth as new feature for a better object localization in the space. Even though we were able to improve the model properties as defined by our probes, it only has a negligible effect on the downstream performance. Our results thus

highlight an important issue of multimodal modeling: the mere presence of information detectable by a probing classifier is not a guarantee that the information is available in a cross-modal setup.

Improving Few-Shot Image Classification Using Machine- and User-Generated Natural Language Descriptions
Kosuke Nishida, Kyosuke Nishida and Shuichi Nishioka 14:15-15:45 (Regency A & B)
Humans can obtain the knowledge of novel visual concepts from language descriptions, and we thus use the few-shot image classification task to investigate whether a machine learning model can have this capability. Our proposed model, LIDE (Learning from Image and DEscription), has a text decoder to generate the descriptions and a text encoder to obtain the text representations of machine- or user-generated descriptions. We confirmed that LIDE with machine-generated descriptions outperformed baseline models. Moreover, the performance was improved further with high-quality user-generated descriptions. The generated descriptions can be viewed as the explanations of the model's predictions, and we observed that such explanations were consistent with prediction results. We also investigated why the language description improves the few-shot image classification performance by comparing the image representations and the text representations in the feature spaces.

FedNLP: Benchmarking Federated Learning Methods for Natural Language Processing Tasks
Bill Yuchen Lin, Chaoyang He, Zihang Ze, Hulin Wang, Yufen Hua, Christophe Dupuy, Rahul Gupta, Mahdi Soltanolkotabi, Xiang Ren and Salman Avestimehr 14:15-15:45 (Regency A & B)
Increasing concerns and regulations about data privacy and sparsity necessitate the study of privacy-preserving, decentralized learning methods for natural language processing (NLP) tasks. Federated learning (FL) provides promising approaches for a large number of clients (e.g., personal devices or organizations) to collaboratively learn a shared global model to benefit all clients while allowing users to keep their data locally. Despite interest in studying FL methods for NLP tasks, a systematic comparison and analysis is lacking in the literature. Herein, we present the FedNLP, a benchmarking framework for evaluating federated learning methods on four different task formulations: text classification, sequence tagging, question answering, and seq2seq. We propose a universal interface between Transformer-based language models (e.g., BERT, BART) and FL methods (e.g., FedAvg, FedOPT, etc.) under various non-IID partitioning strategies. Our extensive experiments with FedNLP provide empirical comparisons between FL methods and help us better understand the inherent challenges of this direction. The comprehensive analysis points to intriguing and exciting future research aimed at developing FL methods for NLP tasks.

Challenging America: Modeling language in longer time scales
Jakub Pokrywka, Filip Graliński, Krzysztof Jassem, Karol Kaczmarek, Krzysztof Jan Jurkiewicz and Piotr Wierzchon 14:15-15:45 (Regency A & B)
The aim of the paper is to apply, for historical texts, the methodology used commonly to solve various NLP tasks defined for contemporary data, i.e. pre-train and fine-tune large Transformer models. This paper introduces an ML challenge, named Challenging America (ChallAm), based on OCR-ed excerpts from historical newspapers collected from the Chronicling America portal. ChallAm provides a dataset of clippings, labeled with metadata on their origin, and paired with their textual contents retrieved by an OCR tool. Three, publicly available, ML tasks are defined in the challenge: to determine the article date, to detect the location of the issue, and to deduce a word in a text gap (cloze test). Strong baselines are provided for all three ChallAm tasks. In particular, we pre-trained a RoBERTa model from scratch from the historical texts. We also discuss the issues of discrimination and hate-speech present in the historical American texts.

PubHealthTab: A Public Health Table-based Dataset for Evidence-based Fact Checking
Mubashara Akhtar, Oana Cocarascu and Elena Simperl 14:15-15:45 (Regency A & B)
Inspired by human fact checkers, who use different types of evidence (e.g. tables, images, audio) in addition to text, several datasets with tabular evidence data have been released in recent years. Whilst the datasets encourage research on table fact-checking, they rely on information from restricted data sources, such as Wikipedia for creating claims and extracting evidence data, making the fact-checking process different from the real-world process used by fact checkers. In this paper, we introduce PubHealthTab, a table fact-checking dataset based on real world public health claims and noisy evidence tables from sources similar to those used by real fact checkers. We outline our approach for collecting evidence data from various websites and present an in-depth analysis of our dataset. Finally, we evaluate state-of-the-art table representation and pre-trained models fine-tuned on our dataset, achieving an overall F_1 score of 0.73.

MM-Claims: A Dataset for Multimodal Claim Detection in Social Media
Gullal Singh Cheema, Sherzod Hakimov, Abdul Sittar, Eric Müller-Budack, Christian Otto and Ralph Ewerth 14:15-15:45 (Regency A & B)
In recent years, the problem of misinformation on the web has become widespread across languages, countries, and various social media platforms. Although there has been much work on automated fake news detection, the role of images and their variety are not well explored. In this paper, we investigate the roles of image and text at an earlier stage of the fake news detection pipeline, called claim detection. For this purpose, we introduce a novel dataset, MM-Claims, which consists of tweets and corresponding images over three topics: COVID-19, Climate Change and broadly Technology. The dataset contains roughly 86000 tweets, out of which 3400 are labeled manually by multiple annotators for the training and evaluation of multimodal models. We describe the dataset in detail, evaluate strong unimodal and multimodal baselines, and analyze the potential and drawbacks of current models.

In-BoXBART: Get Instructions into Biomedical Multi-Task Learning
Mihir Parmar, Swaroop Mishra, Mirali Purohit, Man Luo, Murad Mohammad and Chitta Baral 14:15-15:45 (Regency A & B)
Single-task models have proven pivotal in solving specific tasks; however, they have limitations in real-world applications where multi-tasking is necessary and domain shifts are exhibited. Recently, instructional prompts have shown significant improvement towards multi-task generalization; however, the effect of instructional prompts and Multi-Task Learning (MTL) has not been systematically studied in the biomedical domain. Motivated by this, this paper explores the impact of instructional prompts for biomedical MTL. We introduce the BoX, a collection of 32 instruction tasks for Biomedical NLP across (X) various categories. Using this meta-dataset, we propose a unified model termed as In-BoXBART, that can jointly learn all tasks of the BoX without any task-specific modules. To the best of our knowledge, this is the first attempt to propose a unified model in the biomedical domain and use instructions to achieve generalization across several biomedical tasks. Experimental results indicate that the proposed model: 1) outperforms single-task baseline by 3% and multi-task (without instruction) baseline by 18% on an average, and 2) shows 23% improvement compared to single-task baseline in few-shot learning (i.e., 32 instances per task) on an average. Our analysis indicates that there is significant room for improvement across tasks in the BoX, implying the scope for future research direction.

SemAttack: Natural Textual Attacks via Different Semantic Spaces
Boxin Wang, Chejian Xu, Xiangyu Liu, Yu Cheng and Bo Li 14:15-15:45 (Regency A & B)
Recent studies show that pre-trained language models (LMs) are vulnerable to textual adversarial attacks. However, existing attack methods either suffer from low attack success rates or fail to search efficiently in the exponentially large perturbation space. We propose an efficient and effective framework SemAttack to generate natural adversarial text by constructing different semantic perturbation functions. In particular, SemAttack optimizes the generated perturbations constrained on generic semantic spaces, including typo space, knowledge space (e.g., WordNet), contextualized semantic space (e.g., the embedding space of BERT clusterings), or the combination of these spaces. Thus, the generated adversarial texts are more semantically close to the original inputs. Extensive experiments reveal that state-of-the-art (SOTA) large-scale

LMs (e.g., DeBERTa-v2) and defense strategies (e.g., FreeLB) are still vulnerable to SemAttack. We further demonstrate that SemAttack is general and able to generate natural adversarial texts for different languages (e.g., English and Chinese) with high attack success rates. Human evaluations also confirm that our generated adversarial texts are natural and barely affect human performance. Our code is publicly available at https://github.com/AI-secure/SemAttack.

Language Models for Code-switch Detection of te reo Māori and English in a Low-resource Setting

Jesin James, Vithya Yogarajan, Isabella Shields, Catherine Watson, Peter Keegan, Keoni Mahelona and Peter-Lucas Jones 14:15-15:45 (Regency A & B)

Te reo Māori, New Zealand's only indigenous language, is code-switched with English. Māori speakers are atleast bilingual, and the use of Māori is increasing in New Zealand English. Unfortunately, due to the minimal availability of resources, including digital data, Māori is under-represented in technological advances. Cloud-based multilingual systems such as Google and Microsoft Azure support Māori language detection. However, we provide experimental evidence to show that the accuracy of such systems is low when detecting Māori. Hence, with the support of Māori community, we collect Māori and bilingual data to use natural language processing (NLP) to improve Māori language detection. We train bilingual sub-word embeddings and provide evidence to show that our bilingual embeddings improve overall accuracy compared to the publicly-available monolingual embeddings. This improvement has been verified for various NLP tasks using three bilingual databases containing formal transcripts and informal social media data. We also show that BiLSTM with pre-trained Māori-English sub-word embeddings outperforms large-scale contextual language models such as BERT on down streaming tasks of detecting Māori language. However, this research uses large models 'as is' for transfer learning, where no further training was done on Māori-English data. The best accuracy of 87% was obtained using BiLSTM with bilingual embeddings to detect Māori-English code-switching points.

A Dog Is Passing Over The Jet? A Text-Generation Dataset for Korean Commonsense Reasoning and Evaluation

Jaehyung Seo, Seounghoon Lee, Chanjun Park, Yoonna Jang, Hyeonseok Moon, Sugyeong Eo, Seonmin Koo and Heuiseok Lim 14:15-15:45 (Regency A & B)

Recent natural language understanding (NLU) research on the Korean language has been vigorously maturing with the advancements of pretrained language models and datasets. However, Korean pretrained language models still struggle to generate a short sentence with a given condition based on compositionality and commonsense reasoning (i.e., generative commonsense reasoning). The two major challenges are inadequate data resources to develop generative commonsense reasoning regarding Korean linguistic features and to evaluate language models which are necessary for natural language generation (NLG). To solve these problems, we propose a text-generation dataset for Korean generative commonsense reasoning and language model evaluation. In this work, a semi-automatic dataset construction approach filters out contents inexplicable to commonsense, ascertains quality, and reduces the cost of building the dataset. We also present an in-depth analysis of the generation results of language models with various evaluation metrics along with human-annotated scores. The whole dataset is publicly available at (https://aihub.or.kr/opendata/korea-university).

Restoring Hebrew Diacritics Without a Dictionary

Elazar Gershuni and Yuval Pinter 14:15-15:45 (Regency A & B)

We demonstrate that it is feasible to accurately diacritize Hebrew script without any human-curated resources other than plain diacritized text.

We present Nakdimon, a two-layer character-level LSTM, that performs on par with much more complicated curation-dependent systems, across a diverse array of modern Hebrew sources. The model is accompanied by a training set and a test set, collected from diverse sources.

CoCoA-MT: A Dataset and Benchmark for Contrastive Controlled MT with Application to Formality

Maria Nadejde, Anna Currey, Benjamin Hsu, Xing Niu, Marcello Federico and Georgiana Dinu 14:15-15:45 (Regency A & B)

The machine translation (MT) task is typically formulated as that of returning a single translation for an input segment. However, in many cases, multiple different translations are valid and the appropriate translation may depend on the intended target audience, characteristics of the speaker, or even the relationship between speakers. Specific problems arise when dealing with honorifics, particularly translating from English into languages with formality markers. For example, the sentence "Are you sure?" can be translated in German as "Sind Sie sich sicher?" (formal register) or "Bist du dir sicher?" (informal). Using wrong or inconsistent tone may be perceived as inappropriate or jarring for users of certain cultures and demographics.

This work addresses the problem of learning to control target language attributes, in this case formality, from a small amount of labeled contrastive data. We introduce an annotated dataset (CoCoA-MT) and an associated evaluation metric for training and evaluating formality-controlled MT models for six diverse target languages. We show that we can train formality-controlled models by fine-tuning on labeled contrastive data, achieving high accuracy (82% in-domain and 73% out-of-domain) while maintaining overall quality.

BitextEdit: Automatic Bitext Editing for Improved Low-Resource Machine Translation

Eleftheria Briakou, Sida Wang, Luke Zettlemoyer and Marjan Ghazvininejad 14:15-15:45 (Regency A & B)

Mined bitexts can contain imperfect translations that yield unreliable training signals for Neural Machine Translation (NMT). While filtering such pairs out is known to improve final model quality, we argue that it is suboptimal in low-resource conditions where even mined data can be limited. In our work, we propose instead, to refine the mined bitexts via automatic editing: given a sentence in a language x_f, and a possibly imperfect translation of it $\mathbf{x\_e}$, our model generates a revised version x_f' or x_e' that yields a more equivalent translation pair (i.e., <x_f, x_e'> or <x_f', x_e>). We use a simple editing strategy by (1) mining potentially imperfect translations for each sentence in a given bitext, (2) learning a model to reconstruct the original translations and translate, in a multi-task fashion. Experiments demonstrate that our approach successfully improves the quality of CCMatrix mined bitext for 5 low-resource language-pairs and 10 translation directions by up to 8 BLEU points, in most cases improving upon a competitive translation-based baseline.

Learn To Remember: Transformer with Recurrent Memory for Document-Level Machine Translation

Yukun Feng, Feng Li, Ziang Song, Boyuan Zheng and Philipp Koehn 14:15-15:45 (Regency A & B)

The Transformer architecture has led to significant gains in machine translation. However, most studies focus on only sentence-level translation without considering the context dependency within documents, leading to the inadequacy of document-level coherence. Some recent research tried to mitigate this issue by introducing an additional context encoder or translating with multiple sentences or even the entire document. Such methods may lose the information on the target side or have an increasing computational complexity as documents get longer. To address such problems, we introduce a recurrent memory unit to the vanilla Transformer, which supports the information exchange between the sentence and previous context. The memory unit is recurrently updated by acquiring information from sentences, and passing the aggregated knowledge back to subsequent sentence states. We follow a two-stage training strategy, in which the model is first trained at the sentence level and then finetuned for document-level translation. We conduct experiments on three popular datasets for document-level machine translation and our model has an average improvement of 0.91 s-BLEU over the sentence-level baseline. We also achieve state-of-the-art results on TED and News, outperforming the previous work by 0.36 s-BLEU and 1.49 d-BLEU on average.

Learning to repair: Repairing model output errors after deployment using a dynamic memory of feedback

Niket Tandon, Aman Madaan, Peter Clark and Yiming Yang 14:15-15:45 (Regency A & B)

Large language models (LMs), while powerful, are not immune to mistakes, but can be difficult to retrain. Our goal is for an LM to continue to improve after deployment, without retraining, using feedback from the user. Our approach pairs an LM with (i) a growing memory of cases where the user identified an output error and provided general feedback on how to correct it (ii) a *corrector model*, trained to translate this general feedback into specific edits to repair the model output. Given a new, unseen input, our model can then use feedback from similar, past cases to repair output errors that may occur. We instantiate our approach using an existing, fixed model for *script generation*, that takes a goal (e.g., "bake a cake") and generates a partially ordered sequence of actions to achieve that goal, sometimes containing errors. Our memory-enhanced system learns to apply user feedback to repair such errors (up to 30 points improvement), while making a start at avoiding similar past mistakes on new, unseen examples (up to 7 points improvement in a controlled setting). This is a first step towards strengthening deployed models, potentially broadening their utility. Our code and data is available at `https://github.com/allenai/interscript`

TEAM: A multitask learning based Taxonomy Expansion approach for Attach and Merge
Bornali Phukon, Anasua Mitra, Ranbir Singh Sanasam and Priyankoo Sarmah 14:15-15:45 (Regency A & B)
Taxonomy expansion is a crucial task. Most of Automatic expansion of taxonomy are of two types, attach and merge. In a taxonomy like WordNet, both merge and attach are integral parts of the expansion operations but majority of study consider them separately. This paper proposes a novel mult-task learning-based deep learning method known as Taxonomy Expansion with Attach and Merge (TEAM) that performs both the merge and attach operations. To the best of our knowledge this is the first study which integrates both merge and attach operations in a single model. The proposed models have been evaluated on three separate WordNet taxonomies, viz., Assamese, Bangla, and Hindi. From the various experimental setups, it is shown that TEAM outperforms its state-of-the-art counterparts for attach operation, and also provides highly encouraging performance for the merge operation.

Multimodal Intent Discovery from Livestream Videos
Adyasha Maharana, Quan Hung Tran, Franck Dernoncourt, Seunghyun Yoon, Trung Bui, Walter W Chang and Mohit Bansal 14:15-15:45 (Regency A & B)
Individuals, educational institutions, and businesses are prolific at generating instructional video content such as "how-to" and tutorial guides. While significant progress has been made in basic video understanding tasks, identifying procedural intent within these instructional videos is a challenging and important task that remains unexplored but essential to video summarization, search, and recommendations. This paper introduces the problem of instructional intent identification and extraction from software instructional livestreams. We construct and present a new multimodal dataset consisting of software instructional livestreams and containing manual annotations for both detailed and abstract procedural intent that enable training and evaluation of joint video and text understanding models. We then introduce a multimodal cascaded cross-attention model to efficiently combine the weaker and noisier video signal with the more discriminative text signal. Our experiments show that our proposed model brings significant gains compared to strong baselines, including large-scale pretrained multimodal models. Our analysis further identifies that the task benefits from spatial as well as motion features extracted from videos, and provides insight on how the video signal is preferentially used for intent discovery. We also show that current models struggle to comprehend the nature of abstract intents, revealing important gaps in multimodal understanding and paving the way for future work.

Opponent Modeling in Negotiation Dialogues by Related Data Adaptation
Kushal Chawla, Gale Lucas, Jonathan May and Jonathan Gratch 14:15-15:45 (Regency A & B)
Opponent modeling is the task of inferring another party's mental state within the context of social interactions. In a multi-issue negotiation, it involves inferring the relative importance that the opponent assigns to each issue under discussion, which is crucial for finding high-value deals. A practical model for this task needs to infer these priorities of the opponent on the fly based on partial dialogues as input, without needing additional annotations for training. In this work, we propose a ranker for identifying these priorities from negotiation dialogues. The model takes in a partial dialogue as input and predicts the priority order of the opponent. We further devise ways to adapt related data sources for this task to provide more explicit supervision for incorporating the opponent's preferences and offers, as a proxy to relying on granular utterance-level annotations. We show the utility of our proposed approach through extensive experiments based on two dialogue datasets. We find that the proposed data adaptations lead to strong performance in zero-shot and few-shot scenarios. Moreover, they allow the model to perform better than baselines while accessing fewer utterances from the opponent. We release our code to support future work in this direction.

Learning to Embed Multi-Modal Contexts for Situated Conversational Agents
Haeju Lee, Oh Joon Kwon, Yunseon Choi, Minho Park, Ran Han, Yoonhyung Kim, Jinhyeon Kim, Youngjune Lee, Haebin Shin, Kangwook Lee and Kee-Eung Kim 14:15-15:45 (Regency A & B)
The Situated Interactive Multi-Modal Conversations (SIMMC) 2.0 aims to create virtual shopping assistants that can accept complex multi-modal inputs, i.e. visual appearances of objects and user utterances. It consists of four subtasks, multi-modal disambiguation (MM-Disamb), multi-modal coreference resolution (MM-Coref), multi-modal dialog state tracking (MM-DST), and response retrieval and generation. While many task-oriented dialog systems usually tackle each subtask separately, we propose a jointly learned multi-modal encoder-decoder that incorporates visual inputs and performs all four subtasks at once for efficiency. This approach won the MM-Coref and response retrieval subtasks and nominated runner-up for the remaining subtasks using a single unified model at the 10th Dialog Systems Technology Challenge (DSTC10), setting a high bar for the novel task of multi-modal task-oriented dialog systems.

Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks
Paul Rottger, Bertie Vidgen, Dirk Hovy and Janet B. Pierrehumbert 14:15-15:45 (Regency A & B)
Labelled data is the foundation of most natural language processing tasks. However, labelling data is difficult and there often are diverse valid beliefs about what the correct data labels should be. So far, dataset creators have acknowledged annotator subjectivity, but rarely actively managed it in the annotation process. This has led to partly-subjective datasets that fail to serve a clear downstream use. To address this issue, we propose two contrasting paradigms for data annotation. The descriptive paradigm encourages annotator subjectivity, whereas the prescriptive paradigm discourages it. Descriptive annotation allows for the surveying and modelling of different beliefs, whereas prescriptive annotation enables the training of models that consistently apply one belief. We discuss benefits and challenges in implementing both paradigms, and argue that dataset creators should explicitly aim for one or the other to facilitate the intended use of their dataset. Lastly, we conduct an annotation experiment using hate speech data that illustrates the contrast between the two paradigms.

An Item Response Theory Framework for Persuasion
Anastassia Kornilova, Vladimir Eidelman and Daniel Douglass 14:15-15:45 (Regency A & B)
In this paper, we apply Item Response Theory, popular in education and political science research, to the analysis of argument persuasiveness in language. We empirically evaluate the model's performance on three datasets, including a novel dataset in the area of political advocacy. We show the advantages of separating these components under several style and content representations, including evaluating the ability of the speaker embeddings generated by the model to parallel real-world observations about persuadability.

Self-Supervised Contrastive Learning with Adversarial Perturbations for Defending Word Substitution-based Attacks
Zhao Meng, Yihan Dong, Mrinmaya Sachan and Roger Wattenhofer 14:15-15:45 (Regency A & B)
In this paper, we present an approach to improve the robustness of BERT language models against word substitution-based adversarial attacks

by leveraging adversarial perturbations for self-supervised contrastive learning. We create a word-level adversarial attack generating hard positives on-the-fly as adversarial examples during contrastive learning. In contrast to previous works, our method improves model robustness without using any labeled data. Experimental results show that our method improves robustness of BERT against four different word substitution-based adversarial attacks, and combining our method with adversarial training gives higher robustness than adversarial training alone. As our method improves the robustness of BERT purely with unlabeled data, it opens up the possibility of using large text datasets to train robust language models against word substitution-based adversarial attacks.

The Limits of Word Level Differential Privacy
Justus Mattern, Benjamin Weggenmann and Florian Kerschbaum 14:15-15:45 (Regency A & B)
As the issues of privacy and trust are receiving increasing attention within the research community, various attempts have been made to anonymize textual data. A significant subset of these approaches incorporate differentially private mechanims to perturb word embeddings, thus replacing individual words in a sentence. While these methods represent very important contributions, have various advantages over other techniques and do show anonymization capabilities, they have several shortcomings. In this paper, we investigate these weaknesses and demonstrate significant mathematical constraints diminishing the theoretical privacy guarantee as well as major practical shortcomings with regard to the protection against deanonymization attacks, the preservation of content of the original sentences as well as the quality of the language output. Finally, we propose a new method for text anonymization based on transformer based language models fine-tuned for paraphrasing that circumvents most of the identified weaknesses and also offers a formal privacy guarantee. We evaluate the performance of our method via thourough experimentation and demonstrate superior performance over the discussed mechanisms.

Denoising Neural Network for News Recommendation with Positive and Negative Implicit Feedback
Yunfan Hu, Zhaopeng Qiu and Xian Wu 14:15-15:45 (Regency A & B)
News recommendation is different from movie or e-commercial recommendation as people usually do not grade the news. Therefore, user feedback for news is always implicit (click behavior, reading time, etc). Inevitably, there are noises in implicit feedback. On one hand, the user may exit immediately after clicking the news as he dislikes the news content, leaving the noise in his positive implicit feedback; on the other hand, the user may be recommended multiple interesting news at the same time and only click one of them, producing the noise in his negative implicit feedback. Opposite implicit feedback could construct more integrated user preferences and help each other to minimize the noise influence. Previous works on news recommendation only used positive implicit feedback and suffered from the noise impact. In this paper, we propose a denoising neural network for news recommendation with positive and negative implicit feedback, named DRPN. DRPN utilizes both feedback for recommendation with a module to denoise both positive and negative implicit feedback to further enhance the performance. Experiments on the real-world large-scale dataset demonstrate the state-of-the-art performance of DRPN.

Query2Particles: Knowledge Graph Reasoning with Particle Embeddings
Jiaxin Bai, Zihao Wang, Hongming Zhang and Yangqiu Song 14:15-15:45 (Regency A & B)
Answering complex logical queries on incomplete knowledge graphs (KGs) with missing edges is a fundamental and important task for knowledge graph reasoning. The query embedding method is proposed to answer these queries by jointly encoding queries and entities to the same embedding space. Then the answer entities are selected according to the similarities between the entity embeddings and the query embedding. As the answers to a complex query are obtained from a combination of logical operations over sub-queries, the embeddings of the answer entities may not always follow a uni-modal distribution in the embedding space. Thus, it is challenging to simultaneously retrieve a set of diverse answers from the embedding space using a single and concentrated query representation such as a vector or a hyper-rectangle. To better cope with queries with diversified answers, we propose Query2Particles (Q2P), a complex KG query answering method. Q2P encodes each query into multiple vectors, named particle embeddings. By doing so, the candidate answers can be retrieved from different areas over the embedding space using the maximal similarities between the entity embeddings and any of the particle embeddings. Meanwhile, the corresponding neural logic operations are defined to support its reasoning over arbitrary first-order logic queries. The experiments show that Query2Particles achieves state-of-the-art performance on the complex query answering tasks on FB15k, FB15K-237, and NELL knowledge graphs.

9

Workshops

Overview

During the days of the workshops, **Registration** will be held from 08:00.

Thursday, July 14, 2022

Columbia C	W1 - The 2nd Workshop on Trustworthy NLP	p.193
501 - Chiwawa	W2 - The 3rd Wordplay: When Language Meets Games Workshop	p.194
508 - Tahuya	W3 - The 4th Workshop on Research in Computational Linguistic Typology	p.195
506 - Samish	W4 - The 5th International Workshop on Emoji Understanding and Applications in Social Media	p.197
505 - Queets	W5 - The 6th Workshop on Online Abuse and Harms	p.199
702 - Clearwater	W6 - Clinical NLP 2022: The 4th Clinical Natural Language Processing Workshop	p.200
708 - Sol Duc	W7 - DADC: First Workshop on Dynamic Adversarial Data Collection	p.202
507 - Sauk	W9 - SIGMORPHON 2022: 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology	p.204
502 - Cowlitz	W10 - SUKI: Workshop on Structured and Unstructured Knowledge Integration	p.206
701 - Clallum	W11 - Workshop on Dimensions of Meaning: Distributional and Curated Semantics	p.207
Elwha A	W24 - *SEM 2022: Eleventh Joint Conference on Lexical and Computational Semantics Day 1	p.208
Elwha B	W25 - SemEval-2022: 15th International Workshop on Semantic Evaluation Day 1	p.211
Quinault	W26 - SocialNLP 2022: Tenth International Workshop on Natural Language Processing for Social Media Day 1	p.212

Friday, July 15, 2022

Elwha A	W24 - *SEM 2022: Eleventh Joint Conference on Lexical and Computational Semantics Day 1	p.208
Elwha B	W25 - SemEval-2022: 15th International Workshop on Semantic Evaluation Day 1	p.211
Quinault	W26 - SocialNLP 2022: Tenth International Workshop on Natural Language Processing for Social Media Day 1	p.212
501 - Chiwawa	W12 - BEA-2022: 17th Workshop on Innovative Use of NLP for Building Educational Applications	p.214
508 - Tahuya	W13 - CLPsych: The Eighth Workshop on Computational Linguistics and Clinical Psychology	p.215
Columbia C	W14 - Deep Learning on Graphs for Natural Language Processing	p.216
708 - Sol Duc	W15 - The Fourth Workshop on Narrative Understanding	p.218
507 - Sauk	W19 - MIA: Workshop on Multilingual Information Access	p.219
701 - Clallum	W21 - UnImplicit: The Second Workshop on Understanding Implicit and Underspecified Language	p.220
506 - Samish	W22 - The 3rd Workshop on Automatic Simultaneous Translation	p.221
702 - Clearwater	W23 - The 4th Workshop on Gender Bias in Natural Language Processing	p.222

W1 - The 2nd Workshop on Trustworthy NLP

Organizers:
Apurv Verma, Yada Pruksachatkun, Kai-Wei Chang, Aram Galstyan, Jwala Dhamala, Yang Trista Cao

https://trustnlpworkshop.github.io/
Venue: Columbia C
Thursday, July 14, 2022

Recent progress in Artificial Intelligence (AI) and Natural Language Processing (NLP) has greatly increased their presence in everyday consumer products in the last decade. Common examples include virtual assistants, recommendation systems, and personal healthcare management systems, among others. Advancements in these fields have historically been driven by the goal of improving model performance as measured by accuracy, but recently the NLP research community has started incorporating additional constraints to make sure models are fair and privacy-preserving. However, these constraints are not often considered together, which is important since there are critical questions at the intersection of these constraints such as the tension between simultaneously meeting privacy objectives and fairness objectives, which requires knowledge about the demographics a user belongs to. In this workshop, we aim to bring together these distinct yet closely related topics.

W2 - The 3rd Wordplay: When Language Meets Games Workshop

Organizers:
Prithviraj Ammanabrolu, Xingdi Yuan, Marc-Alexandre Côté, Ashutosh Adhikari, Matthew Hausknecht, Kory Mathewson, Michiaki Tatsubori, Jack Urbanek, Adam Trischler, Jason Weston

`https://wordplay-workshop.github.io/`
Venue: 501 - Chiwawa
Thursday, July 14, 2022

The Wordplay workshop will focus on exploring the utility of interactive narratives, think everything from classic text-adventures like Zork to modern Twine games, to fill a role as the learning environments of choice for language-based tasks including but not limited to storytelling. A few previous iterations of this workshop took place very successfully with hundreds of attendees, at NeurIPS 2018 and NeurIPS 2020. Since then, the community of people working in this area has rapidly increased. This workshop aims to be a centralized place where all researchers involved across a breadth of fields can interact and learn from each other. Furthermore, it will act as a showcase to the wider NLP/RL/Game communities on interactive narrative's place as a learning environment. The program will feature a collection of invited talks in addition to contributed talks and posters from each of these sections of the interactive narrative community and the wider NLP and RL communities.

08:30 - 08:40	***Opening Remarks***
08:40 - 09:25	***Invited Talk 1***
09:25 - 10:00	***Poster Session 1***
10:00 - 10:30	***Break***
10:30 - 11:15	***Invited Talk 2***
11:15 - 12:00	***Invited Talk 3***
12:00 - 13:30	***Lunch***
13:30 - 14:15	***Invited Talk 4***
14:15 - 15:00	***Invited Talk 5***
15:00 - 15:05	***Closing Remarks***
15:05 - 15:30	***Break***
15:30 - 16:00	***Poster Session 2***

W3 - The 4th Workshop on Research in Computational Linguistic Typology

Organizers:
Ekaterina Vylomova, Edoardo Ponti, Ryan Cotterell

`https://www.aclweb.org/portal/content/4th-workshop-research-computational-typology-and-multilingual-nlp`

Venue: 508 - Tahuya

Thursday, July 14, 2022

The aim of the 4th edition of SIGTYP workshop is to act as a platform and a forum for the exchange of information between typology-related research, multilingual NLP, and other research areas that can lead to the development of truly multilingual NLP methods. The workshop is specifically aimed at raising awareness of linguistic typology and its potential in supporting and widening the global reach of multilingual NLP, as well as at introducing computational approaches to linguistic typology. It will foster research and discussion on open problems, not only within the active community working on cross- and multilingual NLP but also inviting input from leading researchers in linguistic typology.

08:30 - 08:40	***Opening Remarks***
08:40 - 09:30	***Grammar Inference for Local Languages: Leveraging Typology for Automatic Grammar Generation (Keynote by Kristen Howell)***
09:30 - 10:00	***Multilingual Representations (Long Talks)***
09:30-09:45	*Multilingualism Encourages Recursion: a Transfer Study with mBERT* Andrea Gregor De Varda and Roberto Zamparelli
09:45-10:00	*Cross-linguistic Comparison of Linguistic Feature Encoding in BERT Models for Typologically Different Languages* Yulia Otmakhova, Karin Verspoor and Jey Han Lau
10:00 - 10:30	***Break***
10:30 - 11:30	***Typology (Short Talks)***
10:30-10:42	*Word-order Typology in Multilingual BERT: A Case Study in Subordinate-Clause Detection* Dmitry Nikolaev and Sebastian Pado
10:42-10:54	*Investigating Information-Theoretic Properties of the Typology of Spatial Demonstratives* Sihan Chen, Richard Futrell and Kyle Mahowald
10:54-11:06	*Tweaking UD Annotations to Investigate the Placement of Determiners, Quantifiers and Numerals in the Noun Phrase* Luigi Talamo
11:06-11:18	*How Universal is Metonymy? Results from a Large-Scale Multilingual Analysis* Temuulen Khishigsuren, Gabor Bella, Thomas Brochhagen, Daariimaa Marav, Fausto Giunchiglia and Khuyagbaatar Batsuren
11:18-11:30	*Typological Word Order Correlations with Logistic Brownian Motion* Kai Hartung, Gerhard Jäger, Sören Gröttrup and Munir Georges
11:30 - 12:00	***Linguistic Trivia***

12:00 - 13:30	***Lunch***
13:30 - 15:00	***Shared Task: Prediction of Cognate Reflexes***
13:30-13:50	*The SIGTYP 2022 Shared Task on the Prediction of Cognate Reflexes* Johann-Mattis List, Ekaterina Vylomova, Robert Forkel, Nathan Hill and Ryan Cotterell
13:50-14:05	*Bayesian Phylogenetic Cognate Prediction* Gerhard Jäger
14:05-14:20	*Mockingbird at the SIGTYP 2022 Shared Task: Two Types of Models for the Prediction of Cognate Reflexes* Christo Kirov, Richard Sproat and Alexander Gutkin
14:20-14:35	*A Transformer Architecture for the Prediction of Cognate Reflexes* Giuseppe Celano
14:35-14:50	*Approaching Reflex Predictions as a Classification Problem Using Extended Phonological Alignments* Tiago Tresoldi
15:00 - 15:30	***Break***
15:30 - 16:20	***Learning from our Differences: How Typologically Distinct Modalities of Data Help Demystify Language Models (Keynote by Isabel Papadimitriou)***
16:20 - 17:00	***Databases and Corpora***
16:20-16:35	*A Database for Modal Semantic Typology* Qingxia Guo, Nathaniel Imel and Shane Steinert-Threlkeld
16:35-16:47	*PaVeDa - Pavia Verbs Database: Challenges and Perspectives* Chiara Zanchi, Silvia Luraghi and Claudia Roberta Combei
16:47-17:00	*ParaNames: A Massively Multilingual Entity Name Corpus* Jonne Sälevä and Constantine Lignos
17:00 - 17:10	***Best Paper Awards, Closing***

W4 - The 5th International Workshop on Emoji Understanding and Applications in Social Media

Organizers:
Sanjaya Wijeratne, Jennifer Lee, Horacio Saggion, Amit Sheth

https://aiisc.ai/emoji2022/
Venue: 506 - Samish
Thursday, July 14, 2022

Pictographs, commonly referred to as "emoji", have become a popular way to enhance electronic communication. With their introduction in the late 1990's, emoji have been widely used to enhance the sentiment, emotion, and sarcasm expressed in social media messages. They often play distinct social and communicative roles compared to other forms of written language while taking over language constructs such as slang terms and emoticons. The ability to automatically process, derive meaning and interpret text fused with emoji will be essential as society embraces emoji as a standard form of online communication. Yet the pictorial nature of emoji, the fact that (the same) emoji may be used in different contexts to express different meanings, and that emoji are used in different cultures and communities over the world who interpret emoji differently, make it especially difficult to apply traditional Natural Language Processing (NLP) techniques to analyze them. To meet these challenges, Emoji aims to stimulate research on understanding social, cultural, communicative, and linguistic roles of emoji and developing novel computational approaches to analyze, interpret and understand emoji and their usage in social media applications. It will provide a forum to bring together researchers and practitioners from both academia and industry in the areas of computer science, natural language processing, computational linguistics, human-computer interaction, and computational social sciences to discuss high-quality research and emerging applications, to exchange ideas and experience, and to identify new opportunities for collaboration.

09:00 - 09:10	***Welcome and Opening Remarks***
09:10 - 10:10	***Keynote - The Next 5000 Years of Emoji Research, Alexander Robertson, Google, USA***
10:10 - 10:30	***Session A - Paper Presentations***
10:10-10:30	*Interpreting Emoji with Emoji* Jens Reelfs, Timon Mohaupt, Sandipan Sikdar, Markus Strohmaier and Oliver Hohlfeld
10:30 - 11:00	***Morning Break***
11:00 - 12:30	***Session B - Paper Presentations***
11:00-11:20	*Understanding the Sarcastic Nature of Emojis with SarcOji* Vandita Grover and Hema Banati
11:20-11:40	*Investigating the Influence of Users Personality on the Ambiguous Emoji Perception* Olga Iarygina
11:40-12:00	*Beyond emojis: an insight into the IKON language* Laura Meloni, Phimolporn Hitmeangsong, Bernhard Appelhaus, Edgar Walthert and Cesco Reale
12:00-12:20	*Graphicon Evolution on the Chinese Social Media Platform BiliBili* Yiqiong Zhang, Susan Herring and Suifu Gan
12:20-12:30	*Conducting Cross-Cultural Research on COVID-19 Memes*

	Jing Ge-Stadnyk and Lusha Sa
12:30 - 14:00	***Lunch Break***
14:00 - 15:00	***Keynote - Processing Emoji in Real-Time, Benjamin Weissman, Rensselaer Polytechnic Institute, USA***
15:00 - 15:45	***Session C - Paper Presentations***
15:00-15:15	*EmojiCloud: a Tool for Emoji Cloud Visualization* Yunhe Feng, Cheng Guo, Bingbing Wen, Peng Sun, Yufei Yue and Dingwen Tao
15:15-15:30	*Semantic Congruency Facilitates Memory for Emojis* Andriana Ge-Christofalos, Laurie Feldman and Heather Sheridan
15:30-15:45	*Emoji semantics/pragmatics: investigating commitment and lying* Benjamin Weissman
15:45 - 15:50	***Closing Remarks***

W5 - The 6th Workshop on Online Abuse and Harms

Organizers:
Kanika Narang, Aida Mostafazadeh Davani, Lambert Mathias, Bertie Vidgen,
Zeerak Talat

https://www.workshoponlineabuse.com/
Venue: 505 - Queets
Thursday, July 14, 2022

For this sixth edition of the Workshop on Online Abuse and Harms (6th WOAH!) we advance research in online abuse through our theme: On Developing Resources and Technologies for low resource Online Abuse and Harms . We continue to emphasize the need for inter-, cross- and anti- disciplinary work on online abuse and harms, and invite paper submissions from a range of fields. These include but are not limited to: NLP, machine learning, computational social sciences, law, politics, psychology, network analysis, sociology and cultural studies. Continuing the tradition started in WOAH 4, we invite civil society, in particular individuals and organisations working with women and marginalised communities who are often disproportionately affected by online abuse, to submit reports, case studies, findings, data, and to record their lived experiences. We hope that through these engagements WOAH can directly address the issues faced by those on the front-lines of tackling online abuse.

08:40 - 08:30	***Welcome + Opening Remarks***
09:25 - 08:40	***Keynote 1 - Mona Diab***
10:10 - 09:25	***Keynote 2 - Murali Shanmugavalan***
10:30 - 10:10	***Morning Break***
11:15 - 10:30	***Keynote 3 - Gebre Gebremeskel***
12:00 - 11:15	***Poster Session (1)***
13:30 - 12:00	***Lunch Break***
14:15 - 13:30	***Poster Session (2)***
15:00 - 14:15	***Keynote 4 - Daniel Borkan, Lucas Dos Santos, Alyssa Lees, and Rachel Rosen***
15:05 - 15:00	***Closing Remarks***

W6 - Clinical NLP 2022: The 4th Clinical Natural Language Processing Workshop

Organizers:
Tristan Naumann, Steven Bethard, Kirk Roberts, Anna Rumshisky

`https://clinical-nlp.github.io/2022/`
Venue: 702 - Clearwater
Thursday, July 14, 2022

Clinical text is growing rapidly as electronic health records become pervasive. Much of the information recorded in a clinical encounter is located exclusively in provider narrative notes, which makes them indispensable for supplementing structured clinical data in order to better understand patient state and care provided. The methods and tools developed for the clinical domain have historically lagged behind the scientific advances in the general-domain NLP. Despite the substantial recent strides in clinical NLP, a substantial gap remains. The goal of this workshop is to address this gap by establishing a regular event in CL conferences that brings together researchers interested in developing state-of-the-art methods for the clinical domain. The focus is on improving NLP technology to enable clinical applications, and specifically, information extraction and modeling of narrative provider notes from electronic health records, patient encounter transcripts, and other clinical narratives.

08:30 - 08:40	***Opening Remarks***
08:40 - 09:25	***Keynote: Mark Dredze***
09:25 - 09:40	***Keynote Q&A***
09:40 - 10:00	***Session 1***
09:40-10:00	*CLPT: A Universal Annotation Scheme and Toolkit for Clinical Language Processing* Saranya Krishnamoorthy, Yanyi Jiang, William Buchanan, Ayush Singh and John E. Ortega
10:00 - 10:20	***Break***
10:20 - 12:00	***Session 2***
10:20-10:40	*PLM-ICD: Automatic ICD Coding with Pretrained Language Models* Chao-Wei Huang, Shang-Chi Tsai and Yun-Nung Chen
10:40-11:00	*An exploratory data analysis: the performance differences of a medical code prediction system on different demographic groups* Heereen Shim, Dietwig Lowet, Stijn Luca and Bart Vanrumste
11:00-11:20	*Cross-Language Transfer of High-Quality Annotations: Combining Neural Machine Translation with Cross-Linguistic Span Alignment to Apply NER to Clinical Texts in a Low-Resource Language* Henning Schäfer, Ahmad Idrissi-Yaghir, Peter A. Horn and Christoph M. Friedrich
11:20-11:40	*Fine-tuning BERT Models for Summarizing German Radiology Findings* Siting Liang, Klaus Kades, Matthias A. Fink, Peter Maximilian Full, Tim Frederik Weber, Jens Kleesiek, Michael Strube and Klaus Maier-Hein
11:40-12:00	*Clinical Flair: A Pre-Trained Language Model for Spanish Clinical Natural Language Processing* Matías Rojas, Jocelyn Dunstan and Fabián Villena
12:00 - 13:30	***Lunch***

13:30 - 15:10	***Session 3***
13:30-13:50	*Exploring Text Representations for Generative Temporal Relation Extraction* Dmitriy Dligach, Steven Bethard, Timothy A Miller and Guergana K Savova
13:50-14:10	*Ensemble-based Fine-Tuning Strategy for Temporal Relation Extraction from the Clinical Narrative* Lijing Wang, Timothy A Miller, Steven Bethard and Guergana K Savova
14:10-14:30	*Learning to Ask Like a Physician* Eric Lehman, Vladislav Lialin, Katelyn Edelwina Yap Legaspi, Anne Janelle R. Sy, Patricia Therese S. Pile, Nicole Rose Alberto, Richard Raymund Reyes Ragasa, Corinna Victoria M. Puyat, Marianne Katharina Vicera Taliño, Isabelle Rose I Alberto, Pia Gabrielle Isidro Alfonso, Dana Moukheiber, Byron C Wallace, Anna Rumshisky, Jennifer J. Liang, Preethi Raghavan, Leo Anthony Celi and Peter Szolovits
14:30-14:50	*What Do You See in this Patient? Behavioral Testing of Clinical NLP Models* Betty Van Aken, Sebastian Herrmann and Alexander Löser
14:50-15:10	*RRED : A Radiology Report Error Detector based on Deep Learning Framework* Dabin Min, Kaeun Kim, Jong Hyuk Lee, Yisak Kim and Chang Min Park
15:10 - 15:30	***Break***
15:30 - 15:50	***Session 4***
15:30-15:50	*m-Networks: Adapting the Triplet Networks for Acronym Disambiguation* Sandaru Seneviratne, Elena Daskalaki, Artem Lenskiy and Hanna Suominen
15:50 - 15:35	***Keynote: Hongfang Liu***
15:35 - 16:50	***Keynote Q&A***
16:50 - 17:00	***Closing Remarks***

W7 - DADC: First Workshop on Dynamic Adversarial Data Collection

Organizers:

Max Bartolo, Hannah Rose Kirk, Pedro Rodriguez, Katerina Margatina, Tristan Thrush, Robin Jia, Pontus Stenetorp, Adina Williams, Douwe Kiela

https://dadcworkshop.github.io/

Venue: 708 - Sol Duc

Thursday, July 14, 2022

Dynamic Adversarial Data Collection (DADC) has been gaining traction in the community as a promising approach to improving data collection practices, model evaluation and performance. DADC allows us to collect human-written data dynamically with models in the loop. Humans can be tasked with finding adversarial examples that fool current state-of-the-art models (SOTA), for example, or they can cooperate with models to find interesting examples. This offers two benefits: it allows us to gauge how good contemporary SOTA methods really are; and it yields data that may be used to train even stronger models by specifically targeting their current weaknesses.

09:00 - 09:10	***Opening Remarks***
09:10 - 09:25	***Collaborative Progress: ML Commons Introduction***
09:25 - 10:00	***Invited Talk 1: Anna Rogers***
10:00 - 10:35	***Invited Talk 2: Jordan Boyd-Graber***
10:35 - 10:50	***Break***
10:50 - 11:10	***Best Paper Talk: Margaret Li and Julian Michael***
11:10 - 11:45	***Invited Talk 3: Sam Bowman***
11:45 - 12:20	***Invited Talk 4: Lora Aroyo***
12:20 - 13:20	***Lunch***
13:20 - 13:55	***Invited Talk 5: Sherry Tongshuang Wu***
13:55 - 14:55	***Panel: The Future of Data Collection***
14:55 - 15:10	***Break***
15:10 - 15:20	***Shared Task Introduction***
15:20 - 15:40	***Shared Task Winners' Presentations***
15:40 - 16:55	***Poster Session***
15:40-15:55	*Resilience of Named Entity Recognition Models under Adversarial Attack* Sudeshna Das and Jiaul Paik
15:55-16:10	*GreaseVision: Rewriting the Rules of the Interface* Siddhartha Datta, Konrad Kollnig and Nigel Shadbolt
16:10-16:25	*Posthoc Verification and the Fallibility of the Ground Truth* Yifan Ding, Nicholas Botzer and Tim Weninger

16:25-16:40	*Overconfidence in the Face of Ambiguity with Adversarial Data* Margaret Li and Julian Michael
16:55 - 17:00	***Closing Session***

W9 - SIGMORPHON 2022: 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology

Organizers:
Garrett Nicolai, Eleanor Chodroff

https://sigmorphon.github.io/workshops/2022/
Venue: 507 - Sauk
Thursday, July 14, 2022

SIGMORPHON aims to bring together researchers interested in applying computational techniques to problems in morphology, phonology, and phonetics. Work that addresses orthographic issues is also welcome. Papers will be on substantial, original, and unpublished research on these topics, potentially including strong work in progress.

08:45 - 09:00	***Opening Remarks***
09:00 - 10:00	***Invited Talk 1: Laura Gwilliams: Parsing continuous speech into lexically bound phonetic sequences***
10:00 - 10:30	***Morning Break***
10:30 - 11:30	***Morning Session: Phonology and Phonetics***
10:30-10:45	*Multidimensional acoustic variation in vowels across English dialects* James Tanner, Morgan Sonderegger and Jane Stuart-Smith
10:45-11:00	*On Building Spoken Language Understanding Systems for Low Resourced Languages* Akshat Gupta
11:00-11:15	*Domain-Informed Probing of wav2vec 2.0 Embeddings for Phonetic Features* Patrick Cormac English, John D. Kelleher and Julie Carson-Berndsen
11:15-11:30	*Investigating phonological theories with crowd-sourced data: The Inventory Size Hypothesis in the light of Lingua Libre* Mathilde Hutin and Marc Allassonnière-Tang
11:30 - 12:30	***Lunch***
12:30 - 13:30	***Invited Talk 2: Gasper Begus: Deep Phonology: Modeling language from raw acoustic data in a fully unsupervised manner***
13:30 - 15:00	***Morning Session: Morphosyntax***
13:30-13:45	*A Masked Segmental Language Model for Unsupervised Natural Language Segmentation* C.M. Downey, Fei Xia, Gina-Anne Levow and Shane Steinert-Threlkeld
13:45-14:00	*Trees probe deeper than strings: an argument from allomorphy* Hossep Dolatian, Shiori Ikawa and Thomas Graf
14:00-14:15	*Logical Transductions for the Typology of Ditransitive Prosody* Mai Ha Vu, Aniello De Santo and Hossep Dolatian
14:15-14:30	*Subword-based Cross-lingual Transfer of Embeddings from Hindi to Marathi and Nepali* Niyata Bafna and Zdeněk Žabokrtský

14:30-14:45	*Morphotactic Modeling in an Open-source Multi-dialectal Arabic Morphological Analyzer and Generator* Nizar Habash, Reham Marzouk, Christian Khairallah and Salam Khalifa
14:45-15:00	*Unsupervised morphological segmentation in a language with reduplication* Simon Todd, Annie Huang, Jeremy Needle, Jennifer Hay and Jeanette King
15:00 - 15:30	***Afternoon Break***
15:30 - 17:45	***Shared Task Session***
17:45 - 18:00	***Closing Statements***

W10 - SUKI: Workshop on Structured and Unstructured Knowledge Integration

Organizers:
Wenhu Chen, Xinyun Chen, Zhiyu Chen, Ziyu Yao, Michihiro Yasunaga, Tao Yu, Rui Zhang

https://suki-workshop.github.io/
Venue: 502 - Cowlitz
Thursday, July 14, 2022

World knowledge is distributed across diverse resources in either structured (tables, lists, graphs, and databases) or unstructured forms (texts, large pretrained language models). Recently, there have been extensive efforts to represent, inject, and ground knowledge in various NLP tasks. Because many downstream applications require the integration of both structured and unstructured knowledge, it is essential to design more generalized models to handle multiple sources of knowledge inputs. However, recent NLP progress is mostly focused on dealing with homogeneous external knowledge resource in a single form. This workshop aims to bring researchers from different backgrounds together to discuss challenges and promote solutions in NLP techniques for jointly dealing with structured and unstructured knowledge. This draws wide attention from multiple NLP areas such as Information Extraction, Question Answering, Semantic Parsing, Information Retrieval, Fact Verification, Summarization, Data-to-Text Generation, and Conversational AI.

09:00 - 08:45	***Opening Remark***
09:45 - 09:00	***Invited Talk by Heng Ji***
10:30 - 09:45	***Invited Talk by Percy Liang***
11:15 - 10:30	***Invited Talk by Jonathan Berant***
12:00 - 11:15	***Invited Talk by Hanna Hajishirzi***
12:30 - 12:00	***Lunch Break***
13:30 - 12:30	***Poster Session***
14:15 - 13:30	***Invited Talk by William Cohen***
15:00 - 14:15	***Invited Talk by Julian Eisenschlos***
16:00 - 15:00	***Shared Task***
16:45 - 16:00	***Invited Talk by Luna Dong***
17:30 - 16:45	***Contributed Talks***
17:45 - 17:30	***Closing Remark***

W11 - Workshop on Dimensions of Meaning: Distributional and Curated Semantics

Organizers:
Collin F. Baker

https://framenet.icsi.berkeley.edu/fndrupal/node/5567/
Venue: 701 - Clallum
Thursday, July 14, 2022

Broadly speaking, computational linguistics research can be divided into two main streams: The first consists of work that relies primarily on operationalizing prior knowledge about language and its use, such as scripts, planning, scenarios, scripts for virtual assistants and FrameNet (FN) frames (Ruppenhofer et al., 2016) as well as lexical databases such as WordNet (Fellbaum 1998), VerbNet (Kipper et al., 2000), and PropBank (Palmer et al., 2005), among others. The second seeks to derive knowledge directly from data (text, speech, and increasingly vision) with unsupervised (or distantly supervised) methods, which are distributional and frequency-based, in Linguistics (Biber et al. 2020), Cognitive Science (Xu and Xu 2021), and Computational Linguistics, notably vector embeddings like BERT (Devlin et al. 2019). They are often complementary: Kuznetsov and Gurevych (2018) combine POS tagging and lemmatization to improve vector embeddings; Qian et al. (2021) combine syntactic knowledge with neural language models to improve accuracy.

13:30 - 13:35	***Welcome and Introduction***
13:35 - 14:25	***Chris Potts (Invited Talk): Lexical semantics in the time of large language models***
13:35 - 15:00	***Session A***
14:25-15:00	*Comparing Distributional and Curated Approaches for Cross-lingual Frame Alignment* Collin F. Baker, Michael Ellsworth, Miriam R. L. Petruck and Arthur Lorenzi
15:00 - 15:30	***Break***
15:30 - 17:00	***Session B***
15:30-16:00	*Multi-sense Language Modelling* Andrea Lekkas, Peter Schneider-Kamp and Isabelle Augenstein
16:00-16:30	*A Descriptive Study of Metaphors and Frames in the Multilingual Shared Annotation Task* Maucha Gamonal
16:30-17:00	*Logical Story Representations via FrameNet + Semantic Parsing* Lane Lawley and Lenhart Schubert

W24 - *SEM 2022: Eleventh Joint Conference on Lexical and Computational Semantics Day 1

Organizers:
Vivi Nastase, Ellie Pavlick, Mohammad Taher Pilehvar, Jose Camacho-Collados, Alessandro Raganato

https://sites.google.com/view/starsem2022/
Venue: Elwha A
Thursday, July 14, 2022 - Friday, July 15, 2022

The 11th Joint Conference on Lexical and Computational Semantics (*SEM 2022) is organized by SIGLEX, the Special Interest Group of the ACL. *SEM brings together researchers interested in the semantics of (many and diverse!) natural languages and its computational modeling. The conference embraces data-driven, neural, and probabilistic approaches, as well as symbolic approaches and everything in between; practical applications as well as theoretical contributions are welcome. The long-term goal of *SEM is to provide a stable forum for the growing number of NLP researchers working on all aspects of semantics of (many and diverse!) natural languages.

08:30 - 10:00 ***Sentence-Level Semantics***

Compositional generalization with a broad-coverage semantic parser
Pia Weißenhorn, Lucia Donatelli and Alexander Koller

DRS Parsing as Sequence Labeling
Minxing Shen and Kilian Evang

Measuring Alignment Bias in Neural Seq2seq Semantic Parsers
Davide Locatelli and Ariadna Quattoni

Semantics-aware Attention Improves Neural Machine Translation
Aviv Slobodkin, Leshem Choshen and Omri Abend

Multilingual Extraction and Categorization of Lexical Collocations with Graph-aware Transformers
Luis Espinosa Anke, Alexander Shvets, Alireza Mohammadshahi, James Henderson and Leo Wanner

Comparison and Combination of Sentence Embeddings Derived from Different Supervision Signals
Hayato Tsukagoshi, Ryohei Sasano and Koichi Takeda

10:00 - 10:30 ***Break***

10:30 - 12:00 ***Evaluation***

Dyna-bAbI: unlocking bAbI's potential with dynamic synthetic benchmarking
Ronen Tamari, Kyle Richardson, Noam Kahlon, Aviad Sar-shalom, Nelson F. Liu, Reut Tsarfaty and Dafna Shahaf

A Dynamic, Interpreted CheckList for Meaning-oriented NLG Metric Evaluation – through the Lens of Semantic Similarity Rating
Laura Zeidler, Juri Opitz and Anette Frank

A Generative Approach for Mitigating Structural Biases in Natural Language Inference
Dimion Asael, Zachary Ziegler and Yonatan Belinkov

How Does Data Corruption Affect Natural Language Understanding Models? A Study on GLUE datasets
Aarne Talman, Marianna Apidianaki, Stergios Chatzikyriakidis and Jörg Tiedemann

AnaLog: Testing Analytical and Deductive Logic Learnability in Language Models
Samuel Ryb, Mario Giulianelli, Arabella Sinclair and Raquel Fernández

Unsupervised Reinforcement Adaptation for Class-Imbalanced Text Classification
Yuexin Wu and Xiaolei Huang

12:00 - 13:30 ***Break***

13:30 - 15:00 ***Invited Talk: Models of meaning? - Jacobs Andreas***

15:00 - 15:30 ***Break***

15:30 - 17:00 ***Lexical Semantics***

Distilling Hypernymy Relations from Language Models: On the Effectiveness of Zero-Shot Taxonomy Induction
Devansh Jain and Luis Espinosa Anke

PropBank Comes of Age—Larger, Smarter, and more Diverse
Sameer Pradhan, Julia Bonn, Skatje Myers, Kathryn Conger, Tim O'gorman, James Gung, Kristin Wright-bettner and Martha Palmer

Pretraining on Interactions for Learning Grounded Affordance Representations
Jack Merullo, Dylan Ebert, Carsten Eickhoff and Ellie Pavlick

Leveraging Three Types of Embeddings from Masked Language Models in Idiom Token Classification
Ryosuke Takahashi, Ryohei Sasano and Koichi Takeda

When Polysemy Matters: Modeling Semantic Categorization with Word Embeddings
Elizabeth Soper and Jean-pierre Koenig

What Drives the Use of Metaphorical Language? Negative Insights from Abstractness, Affect, Discourse Coherence and Contextualized Word Representations
Prisca Piccirilli and Sabine Schulte Im Walde

Assessing the Limits of the Distributional Hypothesis in Semantic Spaces: Trait-based Relational Knowledge and the Impact of Co-occurrences
Mark Anderson and Jose Camacho-collados

08:30 - 10:00 ***Discourse and Dialog***

A Simple Unsupervised Approach for Coreference Resolution using Rule-based Weak Supervision
Alessandro Stolfo, Chris Tanner, Vikram Gupta and Mrinmaya Sachan

Online Coreference Resolution for Dialogue Processing: Improving Mention-Linking on Real-Time Conversations
Liyan Xu and Jinho D. Choi

DeepA2: A Modular Framework for Deep Argument Analysis with Pretrained Neural Text2Text Language Models
Gregor Betz and Kyle Richardson

"What makes a question inquisitive?" A Study on Type-Controlled Inquisitive Question Generation
Lingyu Gao, Debanjan Ghosh and Kevin Gimpel

Speech acts and Communicative Intentions for Urgency Detection
Laurenti Enzo, Bourgon Nils, Farah Benamara, Mari Alda, Véronique Moriceau and Courgeon Camille

What do Large Language Models Learn about Scripts?

Abhilasha Sancheti and Rachel Rudinger

10:00 - 10:30 ***Break***

10:30 - 12:00 ***Events***

Pairwise Representation Learning for Event Coreference
Xiaodong Yu, Wenpeng Yin and Dan Roth

Event Causality Identification via Generation of Important Context Words
Hieu Man, Minh Nguyen and Thien Nguyen

Word-Label Alignment for Event Detection: A New Perspective via Optimal Transport
Amir Pouran Ben Veyseh and Thien Nguyen

Capturing the Content of a Document through Complex Event Identification
Zheng Qi, Elior Sulem, Haoyu Wang, Xiaodong Yu and Dan Roth

Improved Induction of Narrative Chains via Cross-Document Relations
Andrew Blair-stanek and Benjamin Van Durme

12:00 - 13:30 ***Break***

13:30 - 15:00 ***Invited Talk: "Understanding" and prediction: Controlled examinations of meaning sensitivity in pre-trained models - Allyson Ettinger***

15:00 - 15:30 ***Break***

15:30 - 17:00 ***Panel discussion, Closing***

W25 - SemEval-2022: 15th International Workshop on Semantic Evaluation Day 1

Organizers:
Guy Emerson, Natalie Schluter, Gabriel Stanovsky, Ritesh Kumar, Alexis Palmer, Nathan Schneider, Siddharth Singh, Shyam Ratan

`https://semeval.github.io/`
Venue: Elwha B
Thursday, July 14, 2022 - Friday, July 15, 2022

SemEval is a series of international natural language processing (NLP) research workshops whose mission is to advance the current state of the art in semantic analysis and to help create high-quality annotated datasets in a range of increasingly challenging problems in natural language semantics. Each year's workshop features a collection of shared tasks in which computational semantic analysis systems designed by different teams are presented and compared.

08:30 - 10:00	***Invited Talk I: Alane Suhr***
10:00 - 10:30	***Coffee Break***
10:30 - 12:00	***Oral Session I: Task Description Papers***
12:00 - 13:30	***Lunch Break***
13:30 - 15:00	***Poster Session I: System Description Papers***
15:00 - 15:30	***Coffee Break***
15:30 - 17:00	***Poster Session II: System Description Papers***

08:30 - 10:00	***Oral Session II: Task Description Papers***
10:00 - 10:30	***Coffee Break***
10:30 - 12:00	***Poster Session III: System Description Papers***
12:00 - 13:30	***Lunch Break***
13:30 - 15:00	***Invited Talk II: Allyson Ettinger "Understanding" and prediction: Controlled examinations of meaning sensitivity in pre-trained models [Shared with *SEM]***
15:00 - 15:30	***Coffee Break***
15:30 - 16:45	***Poster Session IV: System Description Papers***
16:45 - 17:30	***Oral Session III: Best Paper Awards***

W26 - SocialNLP 2022: Tenth International Workshop on Natural Language Processing for Social Media Day 1

Organizers:
Lun-Wei Ku, Cheng-Te Li, Yu-Che Tsai, Wei-Yao Wang

https://sites.google.com/view/socialnlp2022/
Venue: Quinault
Thursday, July 14, 2022 - Friday, July 15, 2022

With the rapid growth of social networks and Web 2.0 services (e.g. Facebook and Twitter), being able to process data come from such platforms has gained much attention in recent years. SocialNLP is a new inter-disciplinary area of natural language processing (NLP) and social computing. We consider three plausible directions of SocialNLP: (1) addressing issues in social computing using NLP techniques; (2) solving NLP problems using information from social networks or social media; and (3) handling new problems related to both social computing and natural language processing.

08:30 - 08:40	***Opening Remarks***
08:40 - 09:40	***Day-1 Keynote Speech 1 by Prof. Saif Muhammad (National Research Council Canada)***
09:40 - 10:00	***Towards Toxic Positivity Detection***
09:40-10:00	*Towards Toxic Positivity Detection* Ishan Sanjeev Upadhyay, KV Aditya Srivatsa and Radhika Mamidi
10:00 - 10:30	***Day-1 Coffee Break***
10:30 - 11:30	***Day-1 Keynote Speech 2 by Prof. Yohei Seki (University of Tsukuba)***
11:30 - 11:50	***Mask and Regenerate: A Classifier-based Approach for Unpaired Sentiment Transformation of Reviews for Electronic Commerce Websites***
11:30-11:50	*Mask and Regenerate: A Classifier-based Approach for Unpaired Sentiment Transformation of Reviews for Electronic Commerce Websites.* Shuo Yang
11:50 - 13:30	***Day-1 Lunch Break***
13:30 - 14:30	***Day-1 Keynote Speech 3 by Prof. Tim Weninger (University of Notre Dame)***
14:30 - 14:50	***Detecting Rumor Veracity with Only Textual Information by Double-Channel Structure***
14:30-14:50	*Detecting Rumor Veracity with Only Textual Information by Double-Channel Structure* Alex Gunwoo Kim and Sangwon Yoon
14:50 - 15:30	***Day-1 Coffee Break***
15:30 - 16:30	***Day-1 Keynote Speech 4 by Prof. Sonjanya Poria (Singapore University of Technology and Design)***
16:30 - 16:50	***OK Boomer: Probing the socio-demographic Divide in Echo Chambers***
16:30-16:50	*OK Boomer: Probing the socio-demographic Divide in Echo Chambers* Henri-Jacques Geiss, Flora Sakketou and Lucie Flek

08:40 - 09:40	***Day-2 Keynote Speech 1 by Prof. Cristian Danescu-Niculescu-Mizil (Cornell University)***
09:40 - 10:00	***Exploiting Social Media Content for Self-Supervised Style Transfer***
09:40-10:00	*Exploiting Social Media Content for Self-Supervised Style Transfer* Dana Ruiter, Thomas Kleinbauer, Cristina España-Bonet, Josef van Genabith and Dietrich Klakow
10:00 - 10:30	***Day-2 Coffee Break***
10:30 - 11:30	***Day-2 Keynote Speech 2 by Prof. Dan Goldwasser (Purdue University)***
11:30 - 11:50	***Identifying Human Needs through Social Media: A study on Indian cities during COVID-19***
11:30-11:50	*Identifying Human Needs through Social Media: A study on Indian cities during COVID-19* Sunny Rai, Rohan Joseph, Prakruti Singh Thakur and Mohammed Abdul Khaliq
11:50 - 13:30	***Day-2 Lunch Break***
13:30 - 14:30	***Day-2 Keynote Speech 3 by Dr. Ian Stewart (University of Michigan)***
14:30 - 14:50	***A Comparative Study on Word Embeddings and Social NLP Tasks***
14:30-14:50	*A Comparative Study on Word Embeddings and Social NLP Tasks* Fatma Elsafoury, Steven R. Wilson and Naeem Ramzan
14:50 - 15:30	***Day-2 Coffee Break***
15:30 - 16:30	***Day-2 Keynote Speech 4 by Prof. Thamar Solorio (University of Houston)***
16:30 - 16:50	***Leveraging Dependency Grammar for Fine-Grained Offensive Language Detection using Graph Convolutional Networks***
16:30-16:50	*Leveraging Dependency Grammar for Fine-Grained Offensive Language Detection using Graph Convolutional Networks* Divyam Goel and Raksha Sharma
16:50 - 17:00	***Day-2 Closing Remarks***

W12 - BEA-2022: 17th Workshop on Innovative Use of NLP for Building Educational Applications

Organizers:
Ekaterina Kochmar, Jill Burstein, Andrea Horbach, Ronja Laarman-Quante, Nitin Madnani, Anaïs Tack, Victoria Yaneva, Zheng Yuan, Torsten Zesch

https://sig-edu.org/bea/2022
Venue: 501 - Chiwawa
Friday, July 15, 2022

The BEA Workshop is a leading venue for NLP innovation in the context of educational applications. It is one of the largest one-day workshops in the ACL community with over 100 registered attendees in the past several years. The growing interest in educational applications and a diverse community of researchers involved resulted in the creation of the Special Interest Group in Educational Applications (SIGEDU) in 2017, which currently has 240 members.

W13 - CLPsych: The Eighth Workshop on Computational Linguistics and Clinical Psychology

Organizers:
Ayah Zirikly, Dana Atzil-Slonim, Maria Liakata, Steven Bedrick, Bart Desmet, Molly Ireland, Andrew Lee, Sean MacAvaney, Matthew Purver, Rebecca Resnik, Andrew Yates

`https://clpsych.org/`
Venue: 508 - Tahuya
Friday, July 15, 2022

Since 2014, CLPsych has brought together researchers in computational linguistics and NLP, who use computational methods to better understand human language, infer meaning and intention, and predict individuals' characteristics and potential behavior, with mental health practitioners and researchers, who are focused on psychopathology and neurological health and engage directly with the needs of providers and their patients. This workshop's distinctly interdisciplinary nature has improved the exchange of knowledge, fostered collaboration, and increased the visibility of mental health as a problem domain in NLP.

W14 - Deep Learning on Graphs for Natural Language Processing

Organizers:
Lingfei Wu, Bang Liu, Rada Mihalcea, Jian Pei, Yue Zhang, Yunyao Li

https://dlg4nlp-workshop.github.io/dlg4nlp-naacl22/
Venue: Columbia C
Friday, July 15, 2022

There are a rich variety of NLP problems that can be best expressed with graph structures. Due to the great power in modeling non-Euclidean data like graphs or manifolds, deep learning on graphs techniques (i.e., Graph Neural Networks (GNNs)) have opened a new door to solving challenging graph-related NLP problems, and have already achieved great success. As a result, there is a new wave of research at the intersection of deep learning on graphs and NLP which has influenced a variety of NLP tasks, ranging from classification tasks like sentence classification, semantic role labeling, and relation extraction, to generation tasks like machine translation, question generation, and summarization. Despite these successes, deep learning on graphs for NLP still faces many challenges, including but not limited to 1) automatically transforming original text into highly graph-structured data, 2) graph representation learning for complex graphs (e.g., multi-relational graphs, heterogeneous graphs), 3) learning the mapping between complex data structures (e.g., Graph2Seq, Graph2Tree, Graph2Graph).

09:00 - 09:15	***Welcome + Opening Remarks***
09:15 - 10:00	***Keynote Talk 1***
10:00 - 10:45	***Keynote Talk 2***
10:45 - 11:00	***Coffe Break/Social Networking***
11:00 - 11:45	***Panel Discussion***
11:45 - 12:45	***Paper Presentations Session A***
11:45-12:00	*Diversifying Content Generation for Commonsense Reasoning with Mixture of Knowledge Graph Experts* Wenhao Yu, Chenguang Zhu, Lianhui Qin, Zhihan Zhang, Tong Zhao and Meng Jiang
12:00-12:15	*Continuous Temporal Graph Networks for Event-Based Graph Data* Jin Guo, Zhen Han, su Zhou, Jiliang Li, Volker Tresp and Yuyi Wang
12:15-12:30	*Scene Graph Parsing via Abstract Meaning Representation in Pre-trained Language Models* Woo Suk Choi, Yu-Jung Heo, Dharani Punithan and Byoung-Tak Zhang
12:30-12:45	*Explicit Graph Reasoning Fusing Knowledge and Contextual Information for Multi-hop Question Answering* Zhenyun Deng, Yonghua Zhu, Qianqian Qi, Michael Witbrock and Patricia J. Riddle
14:00 - 14:45	***Keynote Talk 3***
14:45 - 15:30	***Keynote Talk 4***
15:30 - 15:45	***Coffe Break/Social Networking***
15:45 - 16:15	***Two Position Talks***

16:15 - 17:15	***Paper Presentations Session B***
16:15-16:30	*Improving Neural Machine Translation with the Abstract Meaning Representation by Combining Graph and Sequence Transformers* Changmao Li and Jeffrey Flanigan
16:30-16:45	*Graph Neural Networks for Adapting Off-the-shelf General Domain Language Models to Low-Resource Specialised Domains* Merieme Bouhandi, Emmanuel Morin and Thierry Hamon
16:45-17:00	*GraDA: Graph Generative Data Augmentation for Commonsense Reasoning* Adyasha Maharana and Mohit Bansal
17:00-17:15	*LiGCN: Label-interpretable Graph Convolutional Networks for Multi-label Text Classification* Irene Li, Aosong Feng, Hao Wu, Tianxiao Li, Toyotaro Suzumura and Ruihai Dong
17:15 - 17:30	***Closing Remarks***

W15 - The Fourth Workshop on Narrative Understanding

Organizers:
Nader Akoury, Faeze Brahman, Khyathi Chandu, Snighda Chaturvedi, Elizabeth Clark, Mohit Iyyer

https://sites.google.com/view/wnu2022/home?authuser=0
Venue: 708 - Sol Duc
Friday, July 15, 2022

This is the 4th iteration of the Narrative Understanding Workshop, which brings together an interdisciplinary group of researchers from AI, ML, NLP, Computer Vision and other related fields, as well as scholars from the humanities to discuss methods to improve automatic narrative understanding capabilities.

08:30 - 09:15	***Virtual Poster Session***
09:15 - 09:20	***Opening Remarks***
09:20 - 10:00	***Invited Talk 1***
10:00 - 10:30	***BREAK***
10:30 - 11:10	***Invited Talk 2***
11:10 - 11:50	***Invited Talk 3***
11:50 - 13:30	***LUNCH***
13:30 - 14:10	***Invited Talk 4***
14:10 - 15:00	***Panel Discussion***
15:00 - 15:30	***BREAK***
15:30 - 16:30	***Hybrid Poster Session***

W19 - MIA: Workshop on Multilingual Information Access

Organizers:
Akari Asai, Eunsol Choi, Jonathan H. Clark, Junjie Hu, Chia-Hsuan Lee, Jungo Kasai, Shayne Lonpre, Ikuya Yamada, Rui Zhang

https://mia-workshop.github.io/
Venue: 507 - Sauk
Friday, July 15, 2022

State-of-the-art NLP technologies such as neural question answering or information retrieval systems have enabled many people to access information efficiently. However, these advances have been made in an English-first way, leaving other languages behind. Large-scale multilingual pre-trained models have achieved significant performance improvements on many multilingual NLP tasks where input text is provided. Yet, on knowledge-intensive tasks that require retrieving knowledge and generating output, we observe limited progress. Moreover, in many languages, existing knowledge sources are critically limited. This workshop addresses challenges for building information access systems in many languages.

09:15 - 09:00	***Opening Remark***
10:15 - 09:15	***Invited talks for the model track***
11:00 - 10:15	***Model panels***
12:00 - 11:00	***Poster session***
13:00 - 12:00	***Lunch break***
13:45 - 13:00	***Shared task session***
14:00 - 13:45	***Best paper talk***
15:00 - 14:00	***Invited talks for the resource track***
15:15 - 15:00	***Break***
16:00 - 15:15	***Resource panels***
16:15 - 16:00	***Closing session***

W21 - UnImplicit: The Second Workshop on Understanding Implicit and Underspecified Language

Organizers:
Valentina Pyatkin, Daniel Fried, Talita Anthonio

https://unimplicit2022.github.io/
Venue: 701 - Clallum
Friday, July 15, 2022

State-of-the-art NLP technologies such as neural question answering or information retrieval systems have enabled many people to access information efficiently. However, these advances have been made in an English-first way, leaving other languages behind. Large-scale multilingual pre-trained models have achieved significant performance improvements on many multilingual NLP tasks where input text is provided. Yet, on knowledge-intensive tasks that require retrieving knowledge and generating output, we observe limited progress. Moreover, in many languages, existing knowledge sources are critically limited. This workshop addresses challenges for building information access systems in many languages.

08:30 - 08:45	***Day-1 Welcome + Opening Remarks***
08:45 - 09:30	***Invited Talk 1***
10:00 - 10:30	***Day-1 Break***
10:30 - 11:00	***Session A***
10:30-10:45	*Pre-trained Language Models' Interpretation of Evaluativity Implicature: Evidence from Gradable Adjectives Usage in Context* Yan Cong
10:45-11:00	*Searching for PETs: Using Distributional and Sentiment-Based Methods to Find Potentially Euphemistic Terms* Patrick Lee, Martha Gavidia, Anna Feldman and Jing Peng
11:00 - 12:00	***Day-1 Poster***
12:00 - 13:30	***Day-1 Lunch***
13:30 - 14:15	***Day-1 Session B***
14:15 - 15:00	***Invited Talk 2***
15:00 - 15:30	***Day-1 Break***
15:30 - 16:15	***Invited Talk 3***
16:15 - 16:45	***Day-1 Session C***
16:45 - 17:00	***Day-1 Closing Remarks***

W22 - The 3rd Workshop on Automatic Simultaneous Translation

Organizers:
Hua Wu, Liang Huang, Zhongjun He, Qun Liu, Wolfgang Macherey, Julia Ive

https://autosimtrans.github.io/
Venue: 506 - Samish
Friday, July 15, 2022

Simultaneous translation, which performs translation concurrently with the source speech, is widely useful in many scenarios such as international conferences, negotiations, press releases, legal proceedings, and medicine. It combines the AI technologies of machine translation (MT), automatic speech recognition (ASR), and text-to-speech synthesis (TTS) and is rapidly becoming a cutting-edge research field. As an emerging and interdisciplinary field, simultaneous translation faces many great challenges.

07:20 - 07:30	***Opening Remarks***
07:30 - 10:30	***Session 1. Invited Talks and Research Paper***
10:10-10:30	*Over-Generation Cannot Be Rewarded: Length-Adaptive Average Lagging for Simultaneous Speech Translation* Sara Papi, Marco Gaido, Matteo Negri and Marco Turchi
10:30 - 18:30	***Break***
18:30 - 20:10	***Session 2. Shared Task***
18:30-18:45	*Findings of the Third Workshop on Automatic Simultaneous Translation* Ruiqing Zhang, Chuanqiang Zhang, Zhongjun He, Hua Wu, Haifeng Wang, Liang Huang, Qun Liu, Julia Ive and Wolfgang Macherey
18:45-19:00	*BIT-Xiaomi's System for AutoSimTrans 2022* Mengge Liu, Xiang Li, Bao Chen, Yanzhi Tian, Tianwei Lan, Silin Li, Yuhang Guo, Jian Luan and Bin Wang
19:00-19:15	*USST's System for AutoSimTrans 2022* Zhu Jia Hui and Yu Jun
19:15-19:30	*System Description on Automatic Simultaneous Translation Workshop* Zecheng Li, Yue Sun and Haoze Li
19:30-19:45	*System Description on Third Automatic Simultaneous Translation Workshop* Zhang Yiqiao
19:45-20:00	*End-to-End Simultaneous Speech Translation with Pretraining and Distillation: Huawei Noah's System for AutoSimTranS 2022* Xingshan Zeng, Pengfei Li, Liangyou Li and Qun Liu

W23 - The 4th Workshop on Gender Bias in Natural Language Processing

Organizers:
Christian Hardmeier, Christine Basta, Marta R. Costa-jussà, Gabriel Stanovsky, Hila Gonen

https://genderbiasnlp.talp.cat/
Venue: 702 - Clearwater
Friday, July 15, 2022

Gender bias, among other demographic biases (e.g. race, nationality, religion), in machine-learned models is of increasing interest to the scientific community and industry. Models of natural language are highly affected by such biases, which are present in widely used products and can lead to poor user experiences. There is a growing body of research into improved representations of gender in NLP models. Key example approaches are to build and use balanced training and evaluation datasets (e.g. Webster et al., 2018), and to change the learning algorithms themselves (e.g. Bolukbasi et al., 2016). While these approaches show promising results, there is more to do to solve identified and future bias issues. In order to make progress as a field, we need to create widespread awareness of bias and a consensus on how to work against it, for instance by developing standard tasks and metrics. Our workshop provides a forum to achieve this goal.

08:30 - 08:40	***Opening Remarks***
08:40 - 09:25	***Keynote 1: Kellie Webster and Kevin Robinson, Google Research***
09:30 - 10:00	***Oral papers 1***
09:30-09:45	*Challenges in Measuring Bias via Open-Ended Language Generation* Afra Feyza Akyürek, Muhammed Yusuf Kocyigit, Sejin Paik and Derry Tanti Wijaya
09:45-10:00	*Looking for a Handsome Carpenter! Debiasing GPT-3 Job Advertisements* Conrad Borchers, Dalia Gala, Benjamin Gilburt, Eduard Oravkin, Wilfried Bounsi, Yuki M Asano and Hannah Kirk
10:00 - 10:30	***Break***
10:30 - 11:15	***Keynote 2: Kai-Wei Chang, UCLA Computer Science***
11:15 - 12:00	***Oral papers 2***
11:15-11:30	*Fewer Errors, but More Stereotypes? The Effect of Model Size on Gender Bias* Yarden Tal, Inbal Magar and Roy Schwartz
11:30-11:45	*Analysis of Gender Bias in Social Perception and Judgement Using Chinese Word Embeddings* Jiali Li, Shucheng Zhu, Ying Liu and Pengyuan Liu
11:45-12:00	*On the Dynamics of Gender Learning in Speech Translation* Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri and Marco Turchi
12:00 - 13:30	***Lunch break***
13:30 - 14:30	***Poster session***
13:30-14:30	*Analyzing Hate Speech Data along Racial, Gender and Intersectional Axes* Antonis Maronikolakis, Philip Baader and Hinrich Schütze

13:30-14:30 *Don’t Forget About Pronouns: Removing Gender Bias in Language Models Without Losing Factual Gender Information*
Tomasz Limisiewicz and David Mareček

13:30-14:30 *Uncertainty and Inclusivity in Gender Bias Annotation: An Annotation Taxonomy and Annotated Datasets of British English Text*
Lucy Havens, Beatrice Alex, Benjamin Bach and Melissa Terras

13:30-14:30 *Debiasing Neural Retrieval via In-batch Balancing Regularization*
Yuantong Li, Xiaokai Wei, Zijian Wang, Shen Wang, Parminder Bhatia, Xiaofei Ma and Andrew Arnold

13:30-14:30 *Gender Biases and Where to Find Them: Exploring Gender Bias in Pre-Trained Transformer-based Language Models Using Movement Pruning*
Przemyslaw Joniak and Akiko Aizawa

13:30-14:30 *Gendered Language in Resumes and its Implications for Algorithmic Bias in Hiring*
Prasanna Parasurama and João Sedoc

13:30-14:30 *The Birth of Bias: A case study on the evolution of gender bias in an English language model*
Oskar Van Der Wal, Jaap Jumelet, Katrin Schulz and Willem Zuidema

13:30-14:30 *Assessing Group-level Gender Bias in Professional Evaluations: The Case of Medical Student End-of-Shift Feedback*
Emmy Liu, Michael Henry Tessler, Nicole Dubosh, Katherine Hiller and Roger Levy

13:30-14:30 *Unsupervised Mitigating Gender Bias by Character Components: A Case Study of Chinese Word Embedding*
Xiuying Chen, Mingzhe Li, Rui Yan, Xin Gao and Xiangliang Zhang

13:30-14:30 *An Empirical Study on the Fairness of Pre-trained Word Embeddings*
Emeralda Sesari, Max Hort and Federica Sarro

13:30-14:30 *Mitigating Gender Stereotypes in Hindi and Marathi*
Neeraja Kirtane and Tanvi Anand

13:30-14:30 *A Taxonomy of Bias-Causing Ambiguities in Machine Translation*
Michal Měchura

13:30-14:30 *On Gender Biases in Offensive Language Classification Models*
Sanjana Marcé and Adam Poliak

13:30-14:30 *Gender Bias in BERT - Measuring and Analysing Biases through Sentiment Rating in a Realistic Downstream Classification Task*
Sophie Jentzsch and Cigdem Turan

13:30-14:30 *Occupational Biases in Norwegian and Multilingual Language Models*
Samia Touileb, Lilja Øvrelid and Erik Velldal

13:30-14:30 *Indigenous Language Revitalization and the Dilemma of Gender Bias*
Oussama Hansal, Ngoc Tan Le and Fatiha Sadat

13:30-14:30 *What changed? Investigating Debiasing Methods using Causal Mediation Analysis*
Sullam Jeoung and Jana Diesner

13:30-14:30 *Why Knowledge Distillation Amplifies Gender Bias and How to Mitigate from the Perspective of DistilBERT*
Jaimeen Ahn, Hwaran Lee, Jinhwa Kim and Alice Oh

13:30-14:30 *Incorporating Subjectivity into Gendered Ambiguous Pronoun (GAP) Resolution using Style Transfer*
Kartikey Pant and Tanvi Dadu

14:30 - 15:00 ***Oral papers 3***

14:30-14:45 *HeteroCorpus: A Corpus for Heteronormative Language Detection*
Juan Vásquez, Gemma Bel-Enguix, Scott Andersen and Sergio-Luis Ojeda-Trueba

14:45-15:00 *Worst of Both Worlds: Biases Compound in Pre-trained Vision-and-Language Models*
Tejas Srinivasan and Yonatan Bisk

15:00 - 15:30 ***Break***

15:30 - 16:15 ***Panel discussion***

16:15 - 16:45 ***Oral papers 4***

16:15-16:30 *Evaluating Gender Bias Transfer from Film Data*
Amanda Bertsch, Ashley Oh, Sanika Natu, Swetha Gangu, Alan W. Black and Emma Strubell

16:45-16:30 *Choose Your Lenses: Flaws in Gender Bias Evaluation*
Hadas Orgad and Yonatan Belinkov

16:30 - 17:00 ***Discussion and closing***

10

Local Guide

Conference Venue

Both as the conference headquarters hotel and the conference venue, the Hyatt Regency Seattle will be your home-away-from-home. Our downtown hotel is walking distance to Seattle's most iconic attractions & experiences. Venture farther & immerse yourself in the Pacific Northwest's incredible landscapes.

Information Regarding Venue Accessibility

The conference and the hotel are committed to providing equal access and opportunity for individuals with disabilities. In addition, the hotel does not operate a shuttle from the airport, but the Light Rail connects to the Westlake station near the hotel; you can read more about Light Rail accessibility here. The hotel has accessible parking in the building itself. The hotel is wheelchair-accessible through the main entrance, the elevators are accessible (auditory or visual feedback or wheelchair height controls, etc.), and bathrooms have accessible stalls that accommodate large electric wheelchairs. There is consistent braille labeling throughout the building. The hotel has gender-inclusive restrooms, lactation rooms, and quiet rooms. The hotel staff has had disability awareness training. Furthermore, all conference events will have automatic or live captioning. The area where the conference will take place is obstacle-free. Service animals are permitted in the hotel. If you would like to request any accommodations (e.g. sign language interpreter, mobility assistance, pregnancy needs, etc.), please reach out to accessibility chairs.
The venue will have a quiet room (703 Hok). It will additionally have a prayer room (704 Newaukum). There will be 2 rooms available for attendees with young children (Rooms 304-305). The venue allows service animals and there will be spaces outdoors for service animals to relieve themselves during the day.

About Seattle

Seattle welcomes millions of visitors from around the world so many of you may already be familiar with all that Seattle has to offer. Sightseeing, arts & culture, cultural heritage, outdoor hiking and enjoying nature, cruising the bay or departing for longer cruises, LGBTQ activities, spectator and other sports, etc. are just some of the things to keep you busy when not at the conference. Check out the Visit Seattle website and read below for ideas and help in planning your Seattle adventures.

Visit Seattle:
`https://visitseattle.org/`
Top 25 Attractions:
`https://visitseattle.org/things-to-do/sightseeing/top-25-attractions/`
Directory of Local Businesses:
`https://visitseattle.org/partners/?type=visitors-guide&design=minority-owned`

Enjoying Seattle

Arts & Culture

Northwest African American Museum

A driving force in keeping the Seattle Black community visible, through events oriented around the celebration of Black leaders that have influenced the region throughout history, and programs/exhibits that highlight local Black voices.

Museum of Museums

Housed in a renovated mid-century medical building, MoM hosts two formal exhibition spaces, three additional on-site museums, rotating installations, murals and sculpture, a theater, weekly art classes, pop-ups, and conceptual gift shop. Nearly every inch of the property is art-activated.

Bainbridge Island Historical Museum

A short and scenic ferry ride away from Downtown Seattle, Bainbridge Island Historical Museum has a number of exhibits on the history of the local area and immigration to the Pacific Northwest.

Frye Art Museum

Reflecting Seattle's evolving identity through exhibitions, programs, and community engagement, the Frye Art Museum showcases local and global artists who are exploring the issues of our time as well as contemporary scholarship on historical subject matter.

Sights

Salish Sea Tours

Native-owned narrated tour of Elliott Bay. Located at Miner's Landing on Pier 57.

Indigenous Walking Tour of the University of Washington Campus

A tour of the picturesque University of Washington (UW) campus with a emphasis on how its development has been influenced by Indigenous cultures.

Ballard Locks (Hiram M. Chittenden Locks)

A popular tourist attraction with a complex of boat locks and the Carl S. English Jr. Botanical Garden. Admission is free, and the locks are often the site of outdoor events like local summer concerts.

Golden Gardens

While Seattle has a number of beaches (notably Alki Beach in West Seattle), one of the most popular is Golden Gardens Park. Swimming is not advised due to the cold water temperatures (Puget Sound averages 66 degrees Fahrenheit), but kitesurfing, kayaking and beach sports are common. There are a number of firepits and picnic areas. The park also has wetlands and hiking trails.

Olympic Sculpture Park

An award-winning park on Seattle's waterfront with a collection of modern sculptures spread out over nine acres. The park is a reclaimed green space - it originated as industrial zone before an effort led by the Seattle Art Museum converted it. Another notable space like this is Gas Works Park in Wallingford, with spectacular views of Downtown Seattle and a "industrial archaeology" site which houses the ruins of a coal gasification plant.

Dining

Annapurna Cafe

Nepalese, Indian and Tibetan restaurant in Capitol Hill. There are excellent dinner options for vegetarians and people with dietary restrictions, and the chai is particularly notable.

Communion

A restaurant/bar in the Central District with rave word-of-mouth reviews that focuses on representing African American culture through food.

Aladdin Falafel Corner and Ugly Mug

Two friendly University District located shops with reasonably priced lunch options. Aladdin is open past midnight for a late dinner as well.

Tea Republik

Friendly neighborhood teahouse in the University District near the University of Washington campus with a large menu of tea options. Ideal for coffee break chats or escaping to get some work done in a quiet corner.

Avole

Ethiopian coffee shop in the Central District with coffee, juice and smoothies that focuses on bringing an Afrofuturism perspective and is inspired by Buna, a community-building Ethiopian coffee-making ceremony.

Rey Amargo

A Mexican chocolate shop in the Capitol Hill neighborhood of Seattle which specializes in artisanal chocolate and coffee beverages, but also has lunch options and churros.

Araya’s Place

An entirely vegan Thai restaurant/bar in the University District with plentiful lunch and dinner options that are likely to appeal to vegans, vegetarians and meat-eaters alike.

Distant Worlds Coffeehouse

A coffee shop with lunch options in the Roosevelt neighborhood that takes pride in being a meeting place and gathering grounds for “[queer, trans, and BIPOC safe geeks] of all stripes.”

Stores & Exhibitions

We highlight a few businesses that are dedicated to honoring, preserving and showcasing the diversity of the Pacific Northwest, particularly Native and queer-friendly businesses.

The Duwamish Longhouse & Cultural Center

4705 W Marginal Way SW, Seattle, WA 98106

Eighth Generation

Pike Place Market

Chief Seattle Club’s Native Works

Pike Place Market

Sacred Circle Gallery

5337 Ballard Ave NW, Seattle, WA 98107

Left Bank Books

Pike Place Market

The Elliot Bay Book Company

1521 10th Ave, Seattle, WA 98122

L.E.M.S. Bookstore

5023 Rainier Ave S, Seattle, WA 98118

Outsider Comics and Geek Boutique

223 N 36th St, Seattle, WA 98103

Useful Information

Here we provide some guidance for getting around in Seattle.

Electricity

The standard US plug is 120V, with a type A/B plug.

Driving

Traffic in the US drives on the right.

Insurance

The Conference Organising Committee or its agents will not be responsible for any medical expenses, loss or accidents incurred during the conference. Delegates are strongly advised to arrange their own personal insurance to cover medical and other expenses including accident or loss. Where a delegate has to cancel for medical reasons, the normal cancellation policy will apply. It is recommended that citizens from EU countries bring with them a current European Health Insurance Card (EHIC) card.

Language

The predominant language is English. The most common secondary languages are Spanish and Chinese.

Money

The US dollar is the currency in Seattle. Circulating coins are commonly in denominations of 1, 5, 10 and 25 cents. Dollars commonly come in denominations of $1, $5, $10, $20, and $50. Foreign exchange services are available throughout the city.

Smoking

Under current legislation, smoking is banned in all indoor public areas and work places, including restaurants and bars. Smoking is still permitted in hotel bedrooms which are designated as smoking bedrooms by the hotel. Smoking in bedrooms in guest houses and bed and breakfast accommodation is at the discretion of the owner.

Tax

Seattle has a 10.25% total sales tax.

Time

Seattle currently operates on Pacific Daylight Time (PDT), which is GMT Greenwich Mean Time - 7 hours.

Shopping

There are shopping areas throughout the city, particularly Downtown Seattle. Founded in 1907, Pike Place Market is a popular spot for tourists with a partially enclosed farmers market, independent shops, restaurants and cafes.

Tipping

Tipping is common, and businesses will often ask for an optional 10-25% tip after a purchase is made.

Weather

Known for its rain, Seattle has a temperate climate, with mild wet winters and relatively cool dry summers. The daily temperature in July in Seattle is on average 72 degrees Fahrenheit. While it tends to stay bright out until late at night in the summer and Seattle rain is generally very mild, it may be wise to bring a light raincoat just in case.

Visa & Passport

Travelers must have a valid passport, or other government-issued ID if a US citizen traveling internally (a NEXUS card is allowed for Canadian citizens). International non-US citizen travelers may also require a nonimmigrant B-1 visa. Visitors are advised to check what form of ID is required by their airline carrier before travelling.

Do I need a visa?

Please see this page (`https://travel.state.gov/content/travel/en/us-visas`) to determine whether you are eligible to apply for an Electronic System Travel Authorization (ESTA) instead, which might be cheaper and easier. This is usually allowed if you hold a passport from a country that participates in the Visa Waiver Program, but there are some exceptions. Be sure to apply for the ESTA at least 72 hours in advance. Do not purchase your ESTA from any website other than the one above.

Covid-19 Safety

Seattle was the first major American city to fully vaccinate 70% of residents 12 and older. King County is now at a vaccination rate of 88% and is safely welcoming back visitors. That said, a specific Covid-19 policy is in place. Conference attendees are expected to wear masks indoors at all times (except during the consumption of food and beverages, or in special cases where a medical exemption has been granted). Hand sanitizer, hot water and soap are available throughout the conference venue for handwashing. To protect fellow attendees, make sure to take a Covid self-test each day before attending events at the conference venue. Please check the conference website and follow emails for the latest information on Covid-19 measures.

Below is a list of available Covid-19 testing locations in Seattle:

1. **Discovery Health Covid-19 Testing**
1209 East Pike St., Seattle, WA 98122
Both Antigen and PCR tests are provided with either same-day or next-day turnaround, prices range from $125-$206 (pricing may change). Please check their website for details and to make appointments.

2. **Curative Covid-19 Testing**
911 Pine St, Seattle, WA 98101 or 705 Pike St, Seattle, WA 98101
Please check their website for pricing and to make an appointment, turnaround time is generally 1-2 days for PCR lab tests and 2 hours for NAAT rapid tests.

3. **GS Labs Covid-19 Testing**
11023 8th Ave NE, Seattle, WA 98125
Antigen tests are provided for $179 and PCR tests for $229 (pricing may change). Appointments can be made on their website.

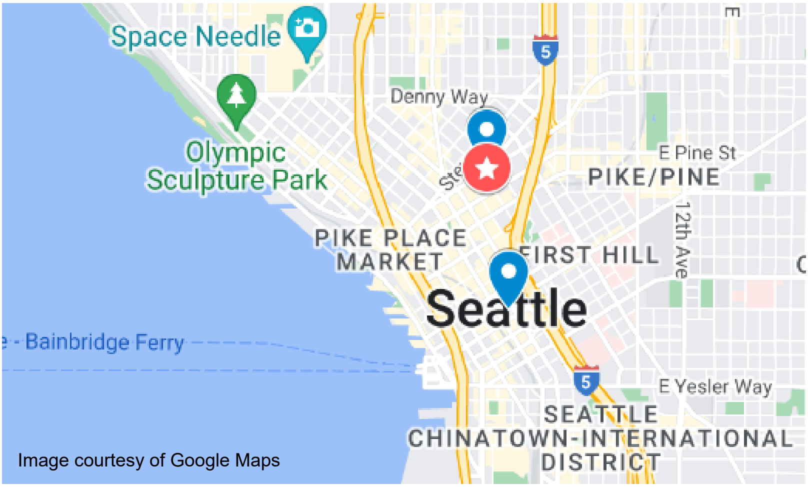
We will also have a point of contact (Yonatan Bisk) that people can report positive cases to and consult with. For any issues or questions regarding NAACL's mask policy, please email the NAACL D&I team (naacl-2022-dei-chairs@googlegroups.com), and for up to date announcements see the website linked below:
`https://2022.naacl.org/covid-19-safety/`.

For other Covid-19 related inquiries while in Seattle, you may contact the Washington State Covid-19 hotline: 1-800-525-0127.

Before you leave, please ensure that you have checked with your airline to understand their policies regarding Covid-19 and have adequate travel insurance in place. You can find this information at `https://travel.state.gov`.

Please ensure you know the Covid-19 testing requirements for return travel back into your country of origin. The requirements are available on your Government website.

Travel to the Conference Venue



By Foot

Downtown Seattle is generally easy to navigate by foot, though visitors should be cautioned that there is steep terrain throughout the area that may pose accessibility challenges.

By Bike

The conference venue will tag and store bicycles for visitors staying at the hotel. Seattle also has bike ride shares (Lime and Veo) for renting electric bikes and scooters via a smartphone app. More information is available on the Seattle.gov website, including information about discounted rates for qualified riders.

By Light Rail

Sound Transit in Seattle runs the extensive Link light rail system, which can provide easy transportation from Seattle-Tacoma (Seatac) Airport to the Northgate area in North Seattle. The closest light rail station to the conference venue is the Westlake Station, which is a short walk from the hotel. Visitors can use a prepaid ORCA smart card as a pass on light rail trains as well as public buses running throughout the city and the King County water taxi. These ORCA cards can be purchased from the Seatac light rail station as well as various other retail locations listed on Sound Transit's website.

By Air

Seattle-Tacoma Airport (Seatac) is located about 14 miles south of Downtown Seattle and is 20 minutes from the conference venue by car. The Seatac light rail station is one of the easiest and most cost-effective ways to travel from the airport to the city center. Alternatively, there is a dedicated area on the 3rd floor of the parking garage for rideshare services like Lyft and Uber. Taxis and car rentals are also available.

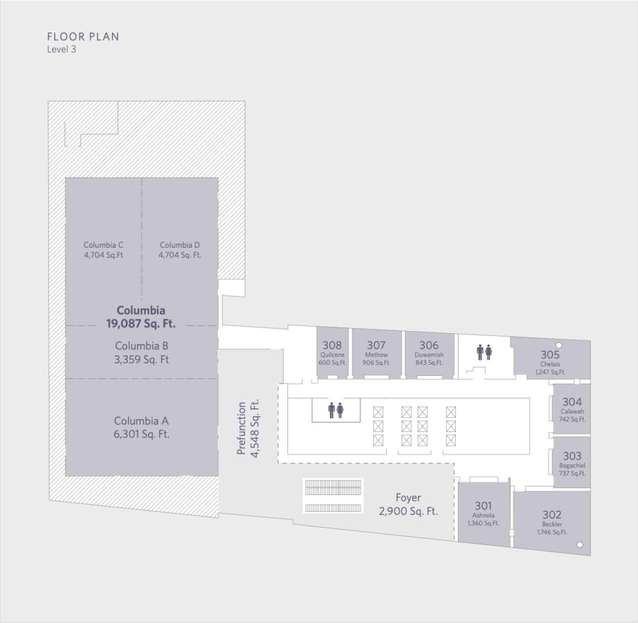
By Car

The conference venue is easily accessible by car using the I-5 highway, which runs through Downtown Seattle. There are several nearby parking garages, including the 808 Howell St. garage which is handicap accessible. There is an hourly rate which varies depending on parking dates and times. The conference venue also has a car rental option.

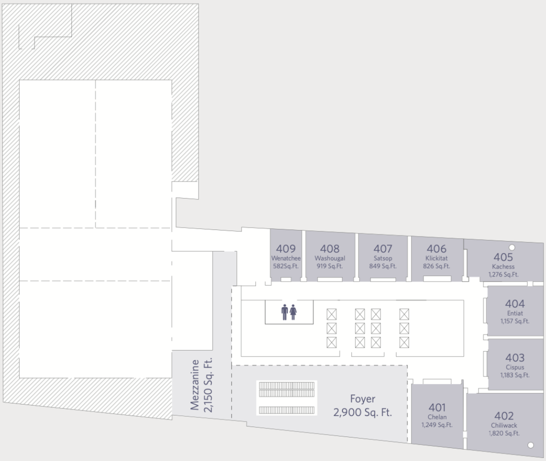
11 Venue Map

Hyatt Regency Seattle Floor Plan

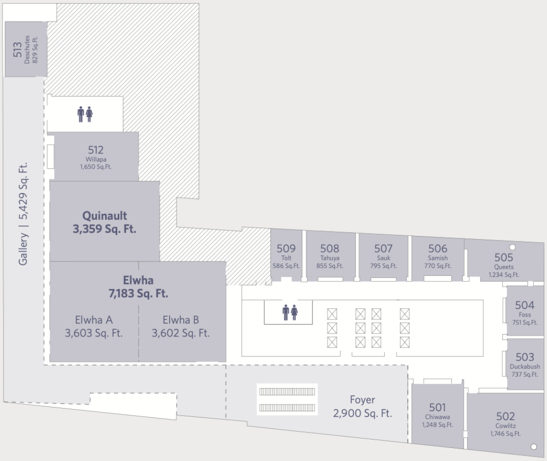
Located in the heart of the Emerald City, Hyatt Regency Seattle is the first and only LEED Gold-Certified hotel in the city. Close to Lake Union and Elliott Bay, and just steps away from dining, shopping, theater, top attractions and both Seattle Convention Center facilities.

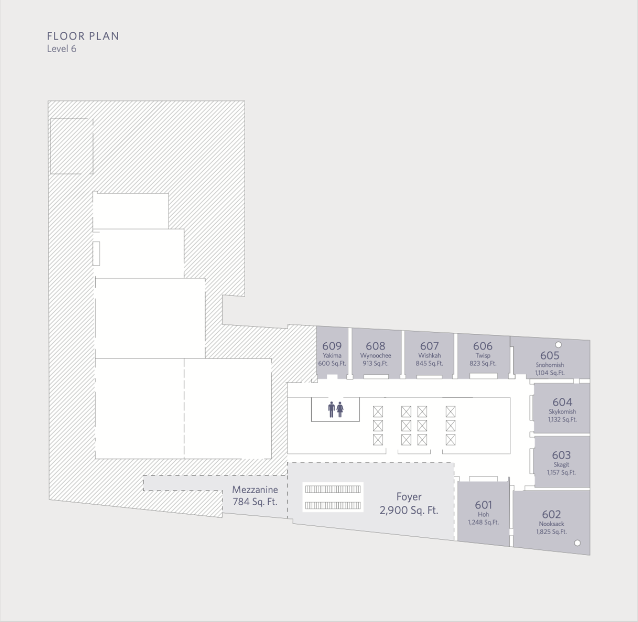


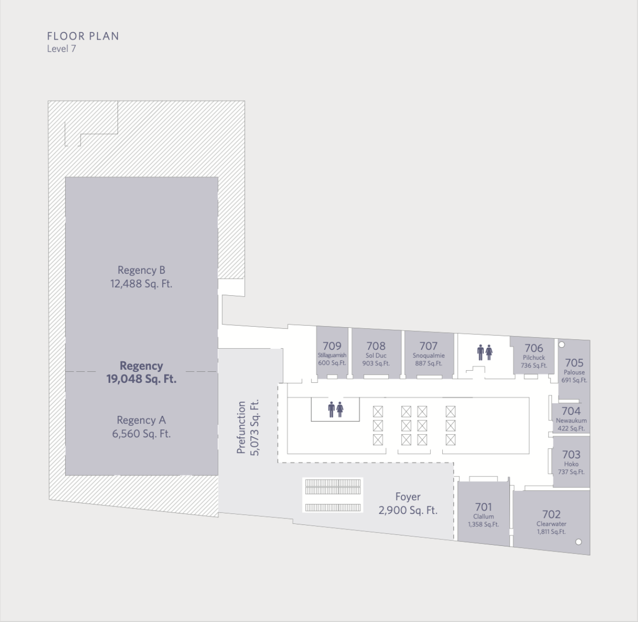
FLOOR PLAN
Level 4



FLOOR PLAN
Level 5




FLOOR PLAN
Level 6
Mezzanine
784 Sq. Ft.
Foyer
2,900 Sq. Ft.
609
Yakima
600 Sq.Ft.
608
Wynoochee
913 Sq.Ft.
607
Wishkah
845 Sq.Ft.
606
Twisp
823 Sq.Ft.
605
Snohomish
1,104 Sq.Ft.
604
Skykomish
1,132 Sq.Ft.
603
Skagit
1,157 Sq.Ft.
601
Hoh
1,248 Sq.Ft.
602
Nooksack
1,825 Sq.Ft.


FLOOR PLAN
Level 7
Regency B
12,488 Sq. Ft.
Regency
19,048 Sq. Ft.
Regency A
6,560 Sq. Ft.
Prefunction
5,073 Sq. Ft.
Foyer
2,900 Sq. Ft.
709
Stillaguamish
600 Sq.Ft.
708
Sol Duc
903 Sq.Ft.
707
Snoqualmie
887 Sq.Ft.
706
Pilchuck
736 Sq.Ft.
705
Palouse
691 Sq.Ft.
704
Newaukum
422 Sq.Ft.
703
Hoko
737 Sq.Ft.
701
Clallum
1,358 Sq.Ft.
702
Clearwater
1,811 Sq.Ft.

Index

amazon | science

Amazon at ACL 2022

Learn more about research at Amazon Science

Amazon Science gives you insight into our approach to customer-obsessed scientific innovation. Our scientists continue to publish, teach, and engage with the academic community, in addition to utilizing our working backwards method to enrich the way we live and work.

To learn more visit:
https://www.amazon.science/

Academics@Amazon

Academics@Amazon are programs aimed at enabling university professors to work on large-scale and high-impact technical challenges at Amazon on a part-time basis without leaving their academic institutions.

Amazon Science Internships

We hire PhD and Masters interns throughout the year across a wide variety of teams, locations, and domains.

Visit: 2022amazonscienceinternships.splashthat.com

Diversity @ Amazon

The ongoing struggles with diversity and inclusion have profoundly influenced the lives and work of scientists and academics around the world. Read their stories, and find out what Amazon's science community is doing to help address these issues.

Job Opportunities

Want to learn more about NLP, speech, ML and other opportunities at @Amazon? Check out our career opportunities today at
http://www.amazon.jobs

Global research locations

Aachen
Atlanta
Austin
Barcelona
Bellevue
Bengaluru
Berkeley
Berlin
Boston
Cambridge
Chennai
Culver City
Cupertino
Dublin
Edinburgh
Gdansk
Graz
Haifa
Hyderabad
Irvine
London
Luxembourg
Arlington
Manhattan Beach
New York
Newark
North Reading
Palo Alto
Pasadena
Pittsburgh
San Diego
San Francisco
Santa Clara
Santa Monica
Seattle
Shanghai
Sunnyvale
Sydney
Tel Aviv
Tokyo
Toronto
Tübingen
Turin
Vancouver
Westborought

Email us directly at acl-2022@amazon.com to learn more.

LIVEPERSON

OUR INDUSTRY-LEADING

Conversational Cloud

Say hello to LivePerson's Conversational Cloud — creating closer connections between brands and customers — all under one roof.

AI and automation capabilities

UNDERSTAND intent from text to automate dynamic actions or routing, conversation automation, chatbots, automated processing, and more.

RESPOND to trends in the marketplace and benchmark against other brands in their vertical via out-of-the-box analytics.

CONNECT contextual data with consumer intent to engage with the right consumer at the right time, for hyper-personalized experiences.

Conversational AI experiences

The tech behind the best brands

We're evolving the tools needed to maximize the performance of machine learning technology so we can get to the future of self-learning Conversational AI chatbots.

Curiously Human™ dialogue

Focused on consumer effort and intent to develop a Curiously Human dialogue, our machine learning Meaningful Automated Conversation Score (MACS) algorithm recognizes when and where the bot fails in the conversation. This provides an additional foundation of self-learning automation to recognize when, where, and how AI chatbots fail, helping improve performance.

High-quality annotation

All of these machine learning tools require annotation, using humans to teach AI models. LivePerson's tools make annotation as easy and scalable as possible — and our annotation team provides the expertise critical to success in solving complex language problems.

Powered by insights and intents from nearly **1 billion conversational interactions every month**, LivePerson's Conversational Cloud delivers exceptional understanding, connection, and business outcome.

VISIT LIVEPERSON.COM TO LEARN MORE

Microsoft Research is where leading scientists and engineers have the freedom and support to propel discovery and innovation. Here, they pursue and publish curiosity-driven research in a range of scientific and technical disciplines that can be translated into products. With access to vast computing power, global multi-disciplinary teams tackle complex problems that drive breakthrough technologies and improve lives.

Careers

Imagine having the freedom and resources to pursue and publish curiosity-driven research that tackles complex problems to improve lives. **aka.ms/msrcareers**

Events

Connect with our researchers at conferences and Microsoft Research events around the world. **aka.ms/msrevent**

Microsoft Research Blog

Read in-depth technical and notable articles from our researchers, scientists, and engineers. **aka.ms/msrblog**

Microsoft Research Podcast

Listen in on conversations that bring you closer to the cutting-edge of technology research and the scientists behind it. **aka.ms/msrpod**

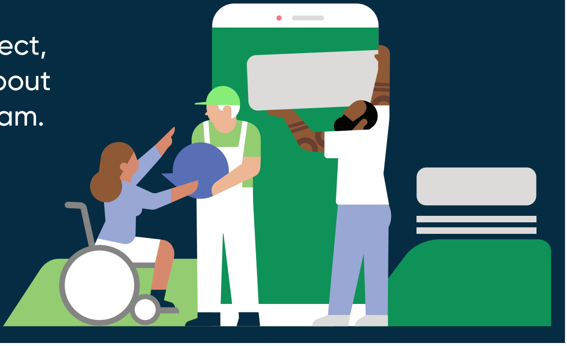
Programs

Further your research with fellowships, grants, and opportunities. **aka.ms/msrprog**

Connect with us:

MicrosoftResearch
@MSFTResearch
microsoftresearch
Microsoft Research Group
@msft_research
#msftresearch

Microsoft

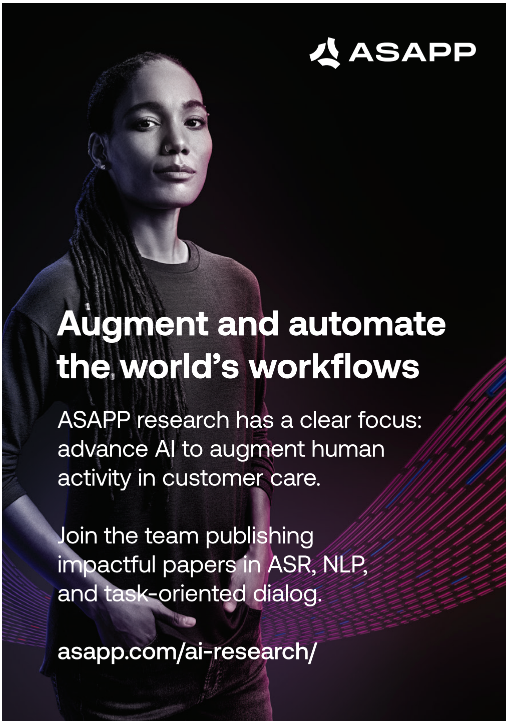


CREATE TODAY.
PREDICT TOMORROW.

SEEK:

“A portrait of a shiba inu astronaut, oil painting, 16th century”

Created with DALL·E, an AI system by OpenAI



TraDeInterpret
myKnowledgeGraph
NLP Pipeline
WordAtlas
Extraggo
Comprehendo
Semantic Paths
LexTag

SPONSORS

DIAMOND

amazon | science

LIVEPERSON

Meta

Google Research

Bloomberg
Engineering

PLATINUM

grammarly

reveal | brainspace

Megagon Labs

ByteDance

Microsoft

GOLD

co:here

Relativity

TWO SIGMA

RESEARCH

OpenAI

TIAA

ETS

servicenow

SILVER

MAGIC DATA
Data Set Your Mind

duolingo

ASAPP

BRONZE

Adobe

human language technology
center of excellence

LinkedIn

babelscape

Rakuten
Institute of Technology

UC SANTA CRUZ
Baskin Engineering

Natural
Language
Processing

D&I CHAMPION

Microsoft

D&I CONTRIBUTOR

RESEARCH