EMNLP 2021

7th - 11th November
Online & in the Dominican Republic

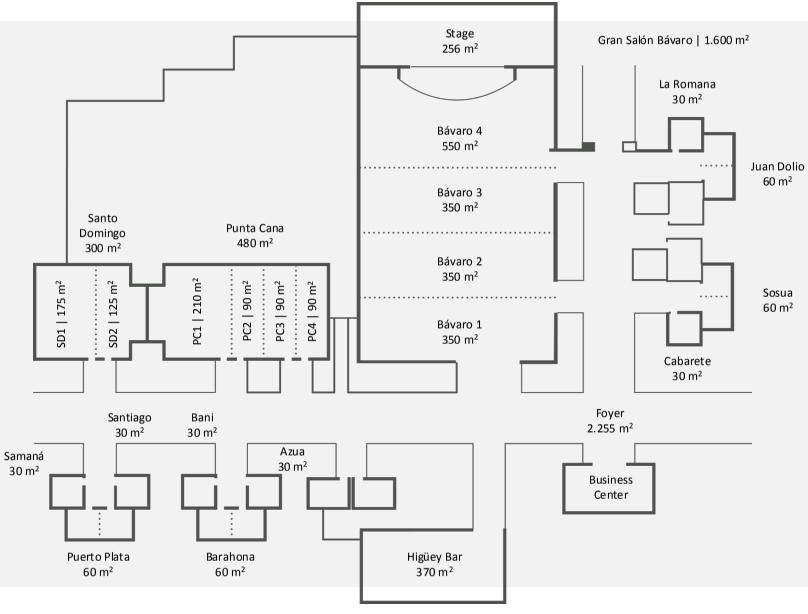
The 2021 Conference on Empirical Methods in Natural Language Processing

---

CONFERENCE HANDBOOK

Meeting rooms

- Themed meetings
- Car exhibitions
- Daylight
- WiFi



Barceló
Bávaro Convention Center

MeetBarceló
HOTEL GROUP

Cover by Nathan Cornille
Handbook assembled by Els Lefever, Pranaydeep Singh and Loic De Langhe
Printing by Omnipress of Madison, Wisconsin

Contents

1

Conference Information

Message from the General Chair

EMNLP 2021 is one of the first hybrid conferences in the field of natural language processing. It is also for us, the organizing team, uncharted domain. Organizing a hybrid conference has felt like organizing two conferences, a virtual one and an in-person one, which seamlessly must work together and with a kind of multi-task objective make the conference experience synergistic and successful both remotely and in person. With this challenge come opportunities. The hybrid format allows remote participation in a conference that is held onsite in Punta Cana, The Dominican Republic, and allows creating a real conference feeling for those who do not want to travel the many miles from the other side of the world and increase their carbon footprint, and for those who have budget restrictions for traveling. We welcome you all!

As in previous years, the purpose of the General Chair's preface is to express thanks to the amazing team of organizing chairs whose heroic efforts made this hybrid conference possible. The organizing team includes:

- The Programme Chairs – Xuanjing Huang, Lucia Specia and Scott Yih – who did a tremendous job to manage the reviewing process and set up an outstanding scientific program.
- The Senior Area Chairs, Area Chairs and Reviewers whose expertise enabled authors to learn from their reviews and to deliver papers that improved on their original submissions.
- The Demonstration Chairs – Heike Adel and Shuming Shi – who selected outstanding demonstrations to complement the program of the main conference.
- The Workshop Chairs – Minlie Huang and Parisa Kordjamshidi – who made a huge effort for organizing hybrid workshops and satellite conferences.
- The Publication Chairs – Loic Barrault, Greg Durrett and Yansong Feng – who met the challenge of identifying and correcting the myriad ways in which papers could be wrongly formatted, and who assembled the result into the conference proceedings.
- The Handbook Chair – Els Lefever – for the timely delivery of handbook information.
- The Publication Chairs of Findings – Gabriel Stanovsky and Tim Van de Cruys – who made it possible that many interesting papers and their findings can be accessed and cited by the public.

- The Tutorial Chairs – Jing Jiang and Ivan Vulic – who selected six excellent tutorials to be presented at the conference.
- The Ethics Chairs – Margot Mieskes and Christopher Potts – who undertook the delicate task of checking papers that had been flagged for potential ethical issues.
- The Website Chair – Miryam de Lhoneux – who ensured that the EMNLP 2021 website promoting this hybrid conference stayed up to date; Mingxiao Li who offered website support; and Nathan Cornille who was responsible for the graphical designs.
- The Virtual Infrastructure Chairs – Quinh Do, Zhaopeng Tu and Dani Yogatama – and the Underline team – Sol Rosenberg, Daniel Luise, Jernej Masnec, Luka Simic, Alexandru Pricop and various support staff.
- The Volunteer Coordinators and Scholarship Chairs – Qi Wu and Diyi Yang – who managed to attract over 200 student and early career volunteers willing to make EMNLP 2021 a success.
- The Publicity Chairs – Raffaella Bernardi and Preethi Jyothi – who have served as both the voice of EMNLP 2021 in communicating with the community and as its ears, reporting on community concerns as soon as they were expressed.
- The Diversity Inclusion Chairs – Laura Alonso Alemany and Toshiaki Nakazawa – who have worked tirelessly to make EMNLP 2021 as welcoming and inclusive as possible for all participants. They have worked with community members to create Birds of a Feather sessions, Affinity Group sessions, student panels and mentoring sessions which contribute to reinforcing the EMNLP community (and sub-groups within this community).

We also want to express special thanks to Priscilla Rasmussen, the ACL Business Manager, first for booking EMNLP 2021 into a beautiful resort in the Dominican Republic, and for the local organization of a hybrid conference. Many thanks, Priscilla!

Finally, we would like to express gratitude to our sponsors, whose generous support has been invaluable in building up EMNLP 2021 to what it is now. These include the Diamond-level sponsors – Apple, Bloomberg Engineering, Facebook AI and Google Research; the Platinum-level sponsors – Amazon Science, Baidu, ByteDance, DeepMind, G Research and Megagon Labs; the Gold-level sponsors – Grammarly and Microsoft; the Silver-level sponsors – duolingo, Naver and Naver Labs Europe; the Bronze-level sponsors – Adobe, Babelscape, human language technology center of excellence and LegalForce; and the Supporter servicenow. We would like to thank the Diversity and Inclusion Champion sponsors – Amazon Science, Deepmind, Google Research and Microsoft; the Diversity and Inclusion Ally sponsor – Morgan Stanley; and the Diversity and Inclusion Contributor sponsors – Adobe and IBM. ACL SIGDAT has also generously contributed to supporting scholarships for attending the conference.

Marie-Francine Moens, KU Leuven, Belgium
EMNLP 2021 General Chair

Message from the Program Committee Co-Chairs

Welcome to the EMNLP 2021, the first hybrid conference in EMNLP's history, which is to be held online and in Punta Cana, Dominican Republic.

EMNLP 2021 has received 3,717 full paper submissions, the largest number to date. After excluding papers withdrawn by the authors, and desk rejecting papers which violated the anonymity policy, the multiple submission policy, or the formatting requirements, we were left with 3,600 submissions to be sent out for review. Despite the record-breaking number of submissions, we were able to keep the acceptance rates at a similar level as past years. 840 submissions were accepted to the main conference. Among them, 315 were accepted as oral papers, and 525 were accepted as posters. The decision between oral and poster presentations was not based on the quality/merit of the papers, but on our understanding of what would be the best format for presentation of each individual paper.

We continued providing the acceptance option of "Findings", following last year's initiative in the form of a companion publication, for papers that narrowly missed acceptance to the main conference, but were judged to be solid, well-executed research, and worthy of publication. After the review process, 445 papers were invited to be included in the Findings. 26 papers declined the offer, leading to 419 papers to be published in the Findings. Some statistics of the accepted papers are shown below.

	Long	Short	Total
Reviewed	2,540	1,060	3,600
Accepted as Oral	249	66	315
Accepted as Poster	401	124	525
Acceptance Rate (Main Conference)	25.6%	17.9%	23.3%
Accepted to Findings	300	119	419
Acceptance Rate (Findings)	11.8%	11.3%	11.6%

To meet the reviewer demands of a large conference, we organized the program committee into 22 tracks, including a special "Multidisciplinary and Area Chair Conflict of Interest" track, based on the track information in past conferences. We also introduced a new track called "Efficient methods for NLP" to promote work aiming to reduce the costs of NLP design and experimentation, similar to the "Green NLP" tracks in EACL 2021 and NAACL 2021. In terms of submissions per track, 9 tracks received more than 200 submissions. Particularly popular were the tracks NLP Applications, Machine Learning for NLP, Machine Translation and Information Extraction, which have around 300 submissions each.

We adopted a hierarchical program committee structure similar to that of recent NLP conferences. For each area, we invited 1-4 Senior Area Chair (SACs), who worked with a team of Area Chairs (ACs) they nominated, as well as an army of reviewers that we put together. We used the submission numbers per track from past conferences to estimate the number of SACs and ACs required for each track, leading to 46 SACs and 236 ACs. For reviewer recruitment, we started with the reviewer lists from past conferences and sent out initial invitations asking reviewers to express their track preferences. We then passed the reviewer list to SACs and asked them to select reviewers from these candidate reviewers based on their expertise, and Semantic/Google Scholar profiles. Overall, this resulted in a total of 3,112 reviewers.

Each submission was assigned to three reviewers and one AC. The initial paper assignment was first made using an automatic algorithm to match the abstracts with ACs/reviewers' past publication records, then adjusted by SACs/PCs. We adapted the review forms from EMNLP 2020, NAACL 2021, and ACL-IJCNLP 2021. Besides the overall recommendation, reviewers were asked to evaluate how reproducible the results in the paper were, and whether there was any ethical concern. Our final decisions were made not just on the review scores, but also took into account the reviews, author responses, discussions among reviewers, meta-reviews and SAC/AC recommendations. To ensure the review quality, we provided detailed guidelines about what

reviewers should and shouldn't do in a review.

We also formed an Ethics Committee (EC) dedicated to ethical issues. 203 papers with ethical concerns raised by the technical reviewing committee were sent to the EC. The EC chairs went over the papers to determine whether a full EC review would be required. If so, the paper received one or two ethics reviews from additional reviewers recruited by the EC chairs. For any paper that was recommended to be accepted based on technical reviews and that had been referred to the EC, the EC chairs recommended one of the following to the PC chairs: (a) accept (12 EMNLP, 11 Findings), (b) conditionally accept (the ethical issues must be addressed in the camera-ready version; 17 EMNLP, 20 Findings), and (c) reject due to ethical issues (1 paper). The authors of all conditionally accepted papers (except 1 paper declining the Findings offer) submitted the camera-ready version and a short response that explained how they had made the changes requested by the EC meta-reviews. The EC chairs double-checked these revised submissions and responses, and confirmed that the ethical concerns had been addressed. As a result, all conditionally accepted papers were accepted to the main conference or Findings.

ACL Rolling Review (ARR) is a new initiative of the Association for Computational Linguistics, where the reviewing and acceptance of papers to publication venues are done in a two-step process: (1) centralized rolling review and (2) submission to a publication venue. Working closely with the ARR organizers, we ran a pilot at EMNLP 2021. 17 papers (16 long, 1 short) were submitted via ARR to EMNLP 2021, accounting for 25% of the ARR May submissions. After the decision process involving only PCs and SACs, 6 papers (5 long, 1 short) were accepted to the main conference, among which 2 papers were accepted orally. The other 5 long papers were accepted to the Findings. These papers will be published in the respective proceedings as any other EMNLP/Findings paper.

Based on the nominations from SACs and ACs, we identified 21 candidates for the best papers and outstanding papers award. These papers are assessed by the Best Paper Award Committee. The award winners will be announced at the closing ceremony.

EMNLP 2021 will also feature 28 papers accepted by the Transactions of the Association for Computational Linguistics (TACL) and 7 papers from the journal of Computational Linguistics (CL), out of which 29 will be presented as orals and 6 as posters.

Another highlight of our program is the three exciting keynote talks, presented by Professor Ido Dagan from Bar-Ilan University, entitled "Where next? Towards multi-text consumption via three inspired research lines", Professor Steven Bird from Charles Darwin University, entitled "LT4All!? Rethinking the Agenda", and Professor Evelina Fedorenko from Massachusetts Institute of Technology, entitled "The language system in the human brain".

There are many people we would like to thank for their significant contributions. EMNLP 2021 would not be possible without their support:

- Our General Chair, Marie-Francine Moens, who has led the whole organizing team, and helped with many of our decision processes;
- 46 SACs who have helped us comprehensively throughout the entire review process, from recruiting ACs and reviewers, assigning papers, checking review quality, making recommendation on final paper decisions, suggesting presentation formats, to recommending best paper candidates; special thanks to Jesse Dodge, who advocated to set up the "Efficient methods for NLP" track, volunteered to serve as the SAC, and helped us update the Reproducibility Checklist to encourage authors to report the computational budget for the experiments in their paper;
- 236 ACs who checked the initial submissions, led paper discussions, wrote meta reviews, ensured review quality, suggested best paper candidates, and recommended outstanding reviewers;
- 3,112 reviewers, 370 secondary reviewers for reviewing papers and actively participating in paper discussions; special thanks to those who stepped in at the last minute to serve as emergency reviewers;

- 35 Ethics Committee members, chaired by Margot Mieskes and Chris Potts, for their hard work to provide ethical reviews and meta-reviews for all papers with serious ethical issues, and ensure that all the conditionally accepted papers have addressed the ethical issues appropriately;

- Best Paper Award Committee: Luke Zettlemoyer (chair), Raffaella Bernardi, Mikel L. Forcada, Pascale Fung, Jianfeng Gao, Min Yen KAN, Heng Ji, Mausam, and Ivan Titov, for selecting best papers and outstanding papers under a tight schedule;

- Our postdoc and student assistants Fernando Alva-Manchego, Zichu Fei, Yiding Tan, Yongxin Zhang and Xingwu Hu, who helped with the initial reviewer assignment, anonymity, multiple submission and format checking;

- Past *ACL PCs, including Trevor Cohn, Yulan He and Yang Liu (EMNLP 2020), Fei Xia, Wenjie Li, Roberto Navigli (ACL-IJCNLP 2021), and Anna Rumshisky, Luke Zettlemoyer and Dilek Hakkani-Tur (NAACL 2021) for all the useful guidance, tips and suggestions on the organization of NLP conferences;

- ARR Editors-in-chief Pascale Fung, Goran Glavaš, Sebastian Riedel, Amanda Stent, and CTO Graham Neubig, for their support in running the first ARR pilot, and providing the code for reviewer COI detection and paper assignment;

- Publication Chairs Loic Barrault, Greg Durrett and Yansong Feng, and Findings Chairs Gabriel Stanovsky and Tim Van de Cruys, for completing the final proceedings within a short period;

- ACL Anthology Director Matt Post, for his help in the production of the conference proceedings;

- TACL editors-in-chief Mark Johnson, Ani Nenkova, and Brian Roark, TACL Editorial Assistant Cindy Robinson, and CL Editor-in-Chief Hwee Tou Ng for coordinating TACL and CL presentations with us;

- Workshop Chairs Parisa Kordjamshidi and Minlie Huang, for connecting Findings paper authors with workshop organizers for possible presentations;

- Publicity Chairs Raffaella Bernardi and Preethi Jyothi, Website Chair Miryam de Lhoneux, and Website Support Mingxiao Li, who announced conference news on EMNLP Website and social media, collected feedback from the community, and disseminated EMNLP papers with potential public interests via media;

- Rich Gerber at SoftConf, who set up the EMNLP conference site, and was always quick to respond to our emails and resolve any problems we encountered with the START system;

- Sol Rosenberg, Daniel Luise and the whole Underline team, for creating the virtual site for the conference and helping put the hybrid program in place;

- Priscilla Rasmussen and members of the Local Organizing Committee, for various discussions on organizing EMNLP, and making the local arrangements for a hybrid programme;

- SIGDAT board members, Iryna Gurevych, Hang Li, Mona Diab and Chin-Yew Lin, for their guidance regarding various decisions;

- 11,425 authors for submitting their work to EMNLP 2021.

Our deepest gratitude to all of you. We hope you will enjoy the hybrid conference experience.

Xuanjing Huang, Fudan University
Lucia Specia, Imperial College London
Scott Wen-tau Yih, Facebook
EMNLP 2021 Program Co-Chairs

Organizing Committee

General Chair
Marie-Francine Moens, KU Leuven

Program Co-chairs
Xuanjing Huang, Fudan University
Lucia Specia, Imperial College London
Scott Wen-tau Yih, Facebook

Workshop Chairs
Parisa Kordjamshidi, Michigan State University
Minlie Huang, Tsinghua University

Tutorial Chairs
Jing Jiang, Singapore Management University
Ivan Vulic, University of Cambridge

Demonstration Chairs
Heike Adel, Bosch Center for Artificial Intelligence
Shuming Shi, Tencent AI lab

Publication Chairs
Loic Barrault, University of Sheffield
Greg Durrett, UT Austin
Yansong Feng, Peking University

Findings Chairs
Gabriel Stanovsky, Hebrew University of Jerusalem
Tim Van de Cruys, KU Leuven

Handbook Chair
Els Lefever, Ghent University
Loic De Langhe, Ghent University
Pranaydeep Singh, Ghent University

Publicity Chairs
Raffaella Bernardi, University of Trento
Preethi Jyothi, IIT Bombay

Website Chair
Miryam de Lhoneux, University of Copenhagen

Website Support
Mingxiao Li, KU Leuven

Virtual Infrastructure Chairs
Zhaopeng Tu, Tencent AI lab
Dani Yogatama, Google DeepMind
Quynh Do, Amazon Aachen

Local Chair
Priscilla Rasmussen, ACL Business Manager

Diversity, Inclusion and Outreach Chairs
Laura Alonso Alemany, Universidad Nacional de Córdoba
Toshiaki Nakazawa, The University of Tokyo

Ethics Committee Chairs
Margot Mieskes, Darmstadt University of Applied Sciences
Christopher Potts, Stanford University

Student Volunteer Coordinator and Scholarship Chairs
Qi Wu, University of Adelaide
Diyi Yang, Georgia Tech
William Held, Georgia Tech

Program Committee

Program Committee Co-chairs
Xuanjing Huang, Fudan University
Lucia Specia, Imperial College London
Scott Wen-tau Yih, Facebook

Area Chairs

Computational Social Science and Cultural Analytics

Jing Jiang, Singapore Management University (Senior Chair)
Alan Ritter, Georgia Institute of Technology (Senior Chair)
Jacob Eisenstein, Google
Kristen Johnson, Michigan State University
David Jurgens, University of Michigan
Preslav Nakov, Qatar Computing Research Institute, HBKU
Simone Paolo Ponzetto, University of Mannheim
Andrew Schwartz, Stony Brook University
Wei Xu, Georgia Institute of Technology
Justine Zhang, Cornell University

Dialogue and Interactive Systems

Annemarie Friedrich, Bosch Center for Artificial Intelligence (Senior Chair)
Kevin Small, Amazon (Senior Chair)
Ruihua Song, Renmin University (Senior Chair)
Emily Dinan, Facebook AI Research
Zhicheng Dou, Renmin University of China
Kallirroi Georgila, University of Southern California
Alborz Geramifard, Facebook AI
Casey Kennington, Boise State University
Yanyan Lan, Tsinghua University
Pearl Pu, EPFL
Oleg Rokhlenko, Amazon Research
Gabriel Skantze, KTH Speech Music and Hearing
Han Ting, National Institute of Advanced Industrial Science and Technology
Stefan Ultes, Mercedes-Benz AG
Thang Vu, University of Stuttgart
Wei Wu, Meituan
Youzheng Wu, JD AI Research
Min Yang, Chinese Academy of Sciences
Sina Zarrieß, University of Bielefeld
Weinan Zhang, Harbin Institute of Technology

Discourse and Pragmatics

Guodong Zhou, Soochow University (Senior Chair)
Fang Kong, Soochow University
Xiaojie Wang, Beijing University of Posts and Telecommunications
Rui Xia, Nanjing University of Science and Technology
Ruifeng Xu, Harbin Institute of Technology, Shenzhen

Efficient Methods for NLP

Jesse Dodge, Allen Institute for AI (Senior Chair)
Gabriel Stanovsky, The Hebrew University of Jerusalem (Senior Chair)
Aleksandr Drozd, RIKEN Center for Computational Science
Jonathan Frankle, MIT

Adhi Kuncoro, University of Oxford and DeepMind
Andreas Rückl, Amazon
Benoît Sagot, Inria
Emma Strubell, Carnegie Mellon University

Ethics in NLP

Natalie Schluter, IT University of Copenhagen (Senior Chair)
Sasha Luccioni, University of Montreal
Marta Ruiz Costa-Jussa, Universitat Politècnica de Catalunya
Simon Suster, University of Melbourne

Generation

Nan Duan, Microsoft Research Asia (Senior Chair)
Mohit Iyyer, University of Massachusetts Amherst (Senior Chair)
Advaith Siddharthan, The Open University (Senior Chair)
Anya Belz, ADAPT Research Centre, Dublin City University
Asli Celikyilmaz, Facebook AI Research
Nina Dethlefs, University of Hull
Angela Fan, Facebook AI Research
Albert Gatt, Utrecht University
Yeyun Gong, Microsoft Research Asia
Meng Jiang, University of Notre Dame
Chenghua Lin, Department of Computer Science, University of Sheffield
Nanyun Peng, University of California, Los Angeles
Hannah Rashkin, Google Research
Hiroya Takamura, The National Institute of Advanced Industrial Science and Technology
Rui Yan, Renmin University of China
Jiajun Zhang, Institute of Automation, Chinese Academy of Sciences

Information Extraction

Leon Derczynski, IT University of Copenhagen (Senior Chair)
Fei Liu, University of Central Florida (Senior Chair)
Kang Liu, Institute of Automation, Chinese Academy of Sciences (Senior Chair)
Christos Christodoulopoulos, Amazon Research
Manuel Ciosici, University of Southern California
Gerard de Melo, Hasso Plattner Institute, University of Potsdam
Antoine Doucet, University of La Rochelle
Doug Downey, Allen Institute for AI, Northwestern University
Mahmoud El-Haj, Lancaster University
Yansong Feng, Peking University
Xianpei Han, Institute of Software, Chinese Academy of Sciences
Luheng He, Google
Zhiyuan Liu, Tsinghua University
Stephen Mayhew, Duolingo
Tim Miller, Boston Children's Hospital and Harvard Medical School
Thien Huu Nguyen, University of Oregon
Qiang Ning, Amazon
Marius Pasca, Google
Hoifung Poon, Microsoft Research
Roi Reichart, Technion - Israel Institute of Technology
Xiang Ren, University of Southern California
Yangqiu Song, HKUST
Jie Yang, Zhejiang University
Dian Yu, Tencent AI Lab
Mo Yu, IBM Research

Information Retrieval and Text Mining

Mark Sanderson, RMIT University (Senior Chair)
Andrew Yates, Max Planck Institute for Informatics (Senior Chair)
Simone Filice, Amazon
Ahmed Hassan Awadallah, Microsoft Research
Evangelos Kanoulas, University of Amsterdam
Sarvnaz Karimi, CSIRO
Heri Ramampiaro, Norwegian University of Science and Technology (NTNU)
Suzan Verberne, LIACS, Leiden University
Thuy Vu, Amazon
Jenny Zhang, RMIT University

Interpretability and Analysis of Models for NLP

Zachary Lipton, Carnegie Mellon University (Senior Chair)
Francesca Toni, Imperial College London (Senior Chair)
Leila Arras, Fraunhofer Heinrich Hertz Institute
Jasmijn Bastings, Google
Grzegorz Chrupała, Tilburg University
Kevin Clark, Google
Oana Cocarascu, King's College London
Yonatan Belinkov, Technion
Sebastian Gehrmann, Google Research
Dieuwke Hupkes, Facebook AI Research
Piyawat Lertvittayakumjorn, Imperial College London
Anna Rogers, University of Copenhagen
Naomi Saphra, New York University
Sameer Singh, University of California, Irvine
Byron Wallace, Northeastern University

Linguistic Theories, Cognitive Modeling and Psycholinguistics

Ryan Cotterell, ETH Zürich (Senior Chair)
Afra Alishahi, Tilburg University
Kyle Mahowald, University of Texas at Austin
Aida Nematzadeh, DeepMind
Adina Williams, Facebook AI

Machine Learning for NLP

Isabelle Augenstein, University of Copenhagen (Senior Chair)
Angeliki Lazaridou, DeepMind (Senior Chair)
Wei Lu, Singapore University of Technology and Design (Senior Chair)
Karthik Narasimhan, Princeton University (Senior Chair)
Yuki Arase, Osaka University
Daniel Beck, University of Melbourne
Iz Beltagy, Allen Institute for AI (AI2)
Kai-Wei Chang, UCLA
Georgiana Dinu, Amazon AWS
Lea Frermann, Melbourne University
Yoav Goldberg, Bar Ilan University
Yoon Kim, MIT, IBM
Carolin Lawrence, NEC Laboratories Europe
Tao Lei, Google
Lei Li, UCSB
André Martins, Unbabel, Instituto de Telecomunicacoes

Vlad Niculae, University of Amsterdam
Siva Reddy, McGill University
Vivek Srikumar, University of Utah
Karl Stratos, Rutgers University
Jun Suzuki, Tohoku University / RIKEN Center for AIP
Swabha Swayamdipta, Allen Institute for Artificial Intelligence
Yulia Tsvetkov, University of Washington
Sam Wiseman, Duke University

Machine Translation and Multilinguality

Alexandra Birch, University of Edinburgh (Senior Chair)
Yang Feng, Institute of Computing Technology, Chinese Academy of Sciences (Senior Chair)
Veselin Stoyanov, Facebook (Senior Chair)
Boxing Chen, Alibaba
Colin Cherry, Google
Trevor Cohn, University of Melbourne
Marcello Federico, Amazon AI
Orhan Firat, Google AI
Dan Garrette, Google Research
Zhongjun He, Baidu, Inc.
Tong Xiao, Northeastern University
Junhui Li, Soochow University, Suzhou
Yang Liu, Tsinghua University
Qun Liu, Huawei Noah's Ark Lab
Rico Sennrich, University of Zurich
Taro Watanabe, Nara Institute of Science and Technology
Deyi Xiong, Tianjin University
Imed Zitouni, Google

NLP Applications

Dina Demner-Fushman, National Library of Medicine (Senior Chair)
Shafiq Rayhan Joty, Nanyang Technological University (Senior Chair)
Maria Liakata, Queen Mary University of London (Senior Chair)
Nikolaos Aletras, University of Sheffield and Amazon
Emilia Apostolova, Language.ai
Steven Bedrick, Oregon Health & Science University
Aoife Cahill, Dataminr
Nancy Chen, Institute for Infocomm Research, A*STAR
Fenia Christopoulou, Huawei Noah's Ark Lab
Nadir Durrani, QCRI
Wei Gao, Singapore Management University
Travis Goodwin, U.S. National Library of Medicine
Yulan He, University of Warwick
Lifu Huang, Virginia Tech
David Mimno, Cornell University
Makoto Miwa, Toyota Technological Institute
Tristan Naumann, Microsoft Research
Dong Nguyen, Utrecht University
Diarmuid Ó Séaghdha, Apple
Nazneen Rajani, Salesforce Research
Kirk Roberts, University of Texas Health Science Center at Houston
Hassan Sajjad, Qatar Computing Research Institute
Yi Tay, Google
Thy Thy Tran, Technische Universität Darmstadt

Karin Verspoor, RMIT University
Chrysoula Zerva, Instituto de Telecomunicações - IST
Aston Zhang, AWS AI
Arkaitz Zubiaga, Queen Mary University of London

Phonology, Morphology and Word Segmentation
Xipeng Qiu, Fudan University (Senior Chair)
Baobao Chang, Institute of Computational Linguistic, Peking University
Ryan Cotterell, ETH Zürich
Hinrich Schütze, University of Munich
Hai Zhao, Shanghai Jiao Tong University

Question Answering
Eunsol Choi, UT Austin (Senior Chair)
Matt Gardner, Allen Institute for Artificial Intelligence (Senior Chair)
Jonathan Berant, Tel Aviv University and AI2
Jordan Boyd-Graber, University of Maryland
Danqi Chen, Princeton University
Yiming Cui, Harbin Institute of Technology
Hannaneh Hajishirzi, University of Washington
Robin Jia, University of Southern California
Tushar Khot, Allen Institute for AI
Tom Kwiatkowski, Google
Kenton Lee, Google Research
Jimmy Lin, University of Waterloo
Minjoon Seo, KAIST
Pontus Stenetorp, University College London
Alon Talmor, Allen Institute for AI, Tel-Aviv University

Resources and Evaluation
Yvette Graham, ADAPT, Trinity College Dublin (Senior Chair)
Markus Freitag, Google Research
Gareth Jones, Dublin City University
Gerasimos Lampouras, Huawei Noah's Ark Lab
Simon Mille, Pompeu Fabra University
Ajay Nagesh, DiDi Labs
Barbara Plank, IT University of Copenhagen
Maja Popović, ADAPT, Dublin City University
Ines Rehbein, University of Mannheim

Semantics: Lexical, Sentence level, Textual Inference and Other areas
Tim Baldwin, The University of Melbourne (Senior Chair)
Sonal Gupta, Facebook (Senior Chair)
James Henderson, Idiap Research Institute (Senior Chair)
Marianna Apidianaki, University of Helsinki
Wai Lam, The Chinese University of Hong Kong
Jey Han Lau, The University of Melbourne
Mike Lewis, Facebook AI Research
Koji Mineshima, Keio University
Nafise Sadat Moosavi, UKP Lab, Technische Universität Darmstadt
Naoaki Okazaki, Tokyo Institute of Technology
Tommaso Pasini, University of Copenhagen
Panupong Pasupat, Google
Michael Roth, University of Stuttgart

Aline Villavicencio, University of Sheffield, UK
Ivan Vulic, University of Cambridge
Diyi Yang, Georgia Institute of Technology
Yi Zhang, Amazon AI

Sentiment Analysis, Stylistic Analysis, and Argument Mining
Veronique Hoste, LT3, Ghent University (Senior Chair)
Yue Zhang, Westlake University (Senior Chair)
Lidong Bing, Alibaba DAMO Academy
Cristina Bosco, Dipartimento di Informatica - Università di Torino
Erik Cambria, Nanyang Technological University
Orphée De Clercq, LT3, Ghent University
Ivan Habernal, Technische Universität Darmstadt
Roman Klinger, University of Stuttgart
Anh Tuan Luu, Nanyang Technological University, Singapore
Soujanya Poria, Singapore University of Technology and Design
Zhiyang Teng, Westlake University
Zhongqing Wang, Soochow University
Zhongyu Wei, School of Data Science, Fudan University
Meishan Zhang, Tianjin University, China

Speech, Vision, Robotics, Multimodal Grounding
Hung-yi Lee, National Taiwan University (Senior Chair)
Pranava Madhyastha, City, University of London (Senior Chair)
Yonatan Bisk, Carnegie Mellon University
Christian Fügen, Facebook AI
David Harwath, The University of Texas at Austin
Lisa Anne Hendricks, DeepMind
Chiori Hori, Mitsubishi Electric Research Laboratories (MERL)
Douwe Kiela, Facebook
Florian Metze, Carnegie Mellon University and Facebook AI
Tara Sainath, Google, Inc.
Radu Soricut, Google LLC
William Wang, University of California, Santa Barbara

Summarization
Xiaojun Wan, Peking University (Senior Chair)
Lu Wang, University of Michigan (Senior Chair)
Giuseppe Carenini, university of british columbia
Michael Elhadad, Ben Gurion University
Pengfei Liu, Carnegie Mellon University
Shashi Narayan, Google
Manabu Okumura, Tokyo Institute of Technology
Jessica Ouyang, University of Texas at Dallas
Maxime Peyrard, EPFL
Caiming Xiong, Salesforce
Rui Zhang, Penn State University

Syntax: Tagging, Chunking and Parsing
Wanxiang Che, Harbin Institute of Technology (Senior Chair)
Liang Huang, Oregon State University and Baidu Research
Zhenghua Li, Soochow University
Weiwei Sun, University of Cambridge
Kewei Tu, ShanghaiTech University

Multidisciplinary and AC COI

Diana Inkpen, University of Ottawa (Senior Chair)
Lluís Màrquez, Amazon SSAI (senior chair)
Cecilia Alm, Rochester Institute of Technology
Paul Cook, University of New Brunswick
Zornitsa Kozareva, Facebook AI
German Rigau, UPV/EHU

Ethics Committee

Margot Mieskes, Darmstadt University of Applied Sciences (Chair)
Christopher Potts, Stanford University (Chair)
George Acquaah-Mensah, Massachusetts College of Pharmacy and Health SCiences
Gilles Adda, LISN-CNRS & LIMSI-CNRS
Hend Al-Khalifa, King Saud University
Hanan Aldarmaki, United Arab Emirates University
Maxime Amblard, Université de Lorraine
Emily M. Bender, University of Washington
Luciana Benotti, Universidad Nacional de Cordoba
Su Lin Blodgett, Microsoft Research
Claudia Borg, University of Malta
Samuel R. Bowman, New York University
Dallas Card, Stanford University
Thiago Castro Ferreira, Federal University of Minas Gerais
Luis Chiruzzo, Universidad de la República
Marie-Catherine de Marneffe, The Ohio State University
Mona Diab, Facebook AI and GWU
Samhaa R. El-Beltagy, Newgiza University/Optomatica
Allyson Ettinger, University of Chicago
Daniel Fried, Facebook AI
Dirk Hovy, Bocconi University
Dan Jurafsky, Stanford University
Min-Yen Kan, National University of Singapore
Roger Levy, Massachusetts Institute of Technology
Margaret Mitchell, Hugging Face
Robert Monarch, Apple
Nyalleng Moorosi, Google AI
Preslav Nakov, Qatar Computing Research Institute, HBKU
Brendan O'Connor, University of Massachusetts Amherst
Jose Ochoa-Luna, Universidad Catolica San Pablo
Ellie Pavlick, Brown University
Manny Rayner, Geneva University
Zeerak Talat, University of Sheffield
Adina Williams, Facebook AI
Jingbo Xia, Huazhong Agricultural University

Welcome Reception

Saturday, November 6, 2021, 6:00pm – 8:00pm

Convention Center Foyer (conference venue)

Get a head start on catching up with your colleagues and registering for the conference. The **Welcome Reception** will take place at the Convention Center Foyer from 18:00 to 20:00 on Saturday, November 6th. A light dinner will be provided.

2

Tutorials: Wednesday, November 10

Overview

9:00 – 12:30 **Morning Tutorials**

Crowdsourcing Beyond Annotation: Case Studies in Benchmark Data Collection
Alane Suhr, Clara Vania, Nikita Nangia, Maarten Sap, Mark Yatskar, Samuel R. Bowman, Yoav Artzi

Financial Opinion Mining
Hsin-Hsi Chen, Hen-Hsen Huang, Chung-Chi Chen

Knowledge-Enriched Natural Language Generation
Wenhao Yu, Meng Jiang, Zhiting Hu, Qingyun Wang, Heng Ji, Nazneen Rajani

Tutorial 1

Crowdsourcing Beyond Annotation: Case Studies in Benchmark Data Collection

Alane Suhr, **Clara Vania**, **Nikita Nangia**, **Maarten Sap**, **Mark Yatskar**, **Samuel R. Bowman**, **Yoav Artzi**

Wednesday, November 10, 2021, 9:00–12:30pm

Crowdsourcing from non-experts is one of the most common approaches to collecting data and annotations in NLP. It has been applied to a plethora of tasks, including question answering, instruction following, visual reasoning, and commonsense reasoning. Even though it is such a fundamental tool, crowdsourcing use is largely guided by common practices and the personal experience of researchers. Developing a theory of crowdsourcing use for practical language problems remains an open challenge. However, there are various principles and practices that have proven effective in generating high quality and diverse data. The goal of this tutorial is to

---

Alane Suhr is a PhD student at Cornell University who's research focuses on grounded natural language understanding. Alane has designed crowdsourcing tasks for collecting language data to study situated natural language understanding. Alane co-presented a tutorial in ACL 2018.

Clara Vania is an applied scientist at Amazon. Her research focuses on crowdsourcing, transfer learning, and multilingual NLU. Recently, she has been working on semi-automatic data collection for natural language inference and crowdsourcing methods for question answering.

Nikita Nangia is a PhD student at New York University. NikitaÕs work focuses on crowdsourcing methods and data creation for natural language understanding. Her recent work explores using incentive structures to illicit creative examples. Nikita co-organized a tutorial on latent structure models for NLP at ACL 2019.

Maarten Sap is a PhD student at the University of Washington. His research focuses on endowing NLP systems with social intelligence and social commonsense, and understanding social inequality and bias in language. His substantial experience with crowdsourcing includes the collecting of the SOCIALIQA commonsense benchmark as well as the creation of knowledge graphs with inferential knowledge (ATOMIC, Social Bias Frames).

Mark Yatskar is an assistant professor at the University of Pennsylvania. His research focuses on the intersection of natural language processing and computer vision. MarkÕs work has resulted in the creation of datasets such as imSitu, QuAC and WinoBias and recent research has focused on gender bias in visual recognition and coreference resolution.

Sam Bowman is an assistant professor at New York University. Sam works on data creation, benchmarking, and model analysis for NLU and computational linguistics. Sam has had a substantial role in several NLU datasets, including SNLI, MNLI, XNLI, CoLA, and BLiMP, and his recent work has focused on experimentally evaluating methods for crowdsourced corpus construction.

Yoav Artzi is an associate professor at Cornell University. YoavÕs research focuses on learning expressive models for natural language understanding, most recently in situated interactive scenarios. Yoav led tutorials on semantic parsing in ACL 2013, EMNLP 2014 and AAAI 2015.

expose NLP researchers to such data collection crowdsourcing methods and principles through a detailed discussion of a diverse set of case studies.

Tutorial 2

Financial Opinion Mining

Hsin-Hsi Chen, **Hen-Hsen Huang**, **Chung-Chi Chen**

Wednesday, November 10, 2021, 9:00–12:30pm

In this tutorial, we disassemble a financial opinion into 12 components. This tutorial starts by introducing the components one by one and introduces the related studies from both NLP technical aspects and the real-world applications. Besides, in the FinTech trend, financial service gets much attention from the financial industry. However, few studies discuss the opinion toward financial service. In this tutorial, we will also introduce this kind of opinion and provide a comparison with the opinion of investors and customer's opinions in other industries. Several unexplored research questions will be proposed. The audiences of this tutorial will gain

---

Chung-Chi Chen is a postdoctoral researcher at the MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan. He got the Ph.D. degree in the Department of Computer Science and Information Engineering at Na- tional Taiwan University. He received the M.S. degree in Quantitative Finance from National Tsing Hua University, Taiwan. His research focuses on opinion mining and sentiment analysis in finance. He is the organizer of FinNum shared task series in NTCIR (2018-2022) and the FinNLP workshop series in IJCAI (2019-2021). He is the presenter of the AACL-2020 "Natural Language Processing in Financial Technology Applications" tutorial and the presenter of the EMNLP-2021 "Financial Opinion Mining" tutorial. His work has been published in ACL, WWW, SIGIR, IJCAI, and CIKM, and served as PC members in ACL, AAAI, EMNLP, CIKM, and WSDM. He won the 1st prize in both the Jih Sun FinTech Hackathon (2019) and the Standard Chartered FinTech competition (2018), and the 2nd prize in both the Jih Sun FinTech Hackathon (2018) and the E.SUN FHC FinTech Hackathon (2017).

Hen-Hsen Huang is an assistant research fellow at the Institue of Information Science, Academia Sinica, Taiwan. His research interests include natural language processing and information retrieval. His work has been published in ACL, SI- GIR, WWW, IJCAI, CIKM, COLING, and so on. Dr. Huang received the Honorable Mention of Doctoral Dissertation Award of ACLCLP in 2014 and the Honorable Mention of Master Thesis Award of ACLCLP in 2008. He served as the registration chair of TAAI 2017, the publication chair of ROCLING 2020, and as PC members of representative conferences in computational linguistics including ACL, COLING, EMNLP, and NAACL. He was one of organizers of FinNum Task at NTCIR and FinNLP Workshop at IJCAI.

Hsin-Hsi Chen is a professor in the Department of Computer Science and Information Engineering, National Taiwan University. He was conference chair of IJCNLP 2013, program co-chair of ACM SIGIR 2010, senior PC member of ACM SIGIR 2006, 2007, 2008 and 2009, area/track chair of AAAI 2020, EMNLP 2018, ACL 2012, ACL-IJCNLP 2009 and ACM CIKM 2008, and PC member of many conferences (IJCAI, SIGIR, WSDM, ACL, COLING, EMNLP, NAACL, EACL, IJCNLP, WWW, and so on). He will be conference chair of ACM SIGIR 2023. He received Google research awards in 2007 and 2012, MOST Outstanding Research Award in 2017, and the AmTRAN Chair Professorship in 2018.

an overview of financial opinion mining and figure out their research directions based on the proposed research agenda.

Tutorial 3

Knowledge-Enriched Natural Language Generation

Wenhao Yu, **Meng Jiang**, **Zhiting Hu**, **Qingyun Wang**, **Heng Ji**, **Nazneen Rajani**

Wednesday, November 10, 2021, 9:00–12:30pm

Knowledge-enriched text generation poses unique challenges in modeling and learning, driving active research in several core directions, ranging from integrated modeling of neural representations and symbolic information in the sequential/hierarchical/graphical structures, learning without direct supervisions due to the cost of structured annotation, efficient optimization and inference with massive and global constraints, to language grounding on multiple modalities, and generative reasoning with implicit commonsense knowledge and background knowledge. In this tutorial we will present a roadmap to line up the state-of-the-art methods to tackle these challenges on this cutting-edge problem. We will dive deep into various technical components: how to represent knowledge, how to feed knowledge into a generation model, how to evaluate generation results, and what are the remaining challenges?

Wenhao Yu is a Ph.D. student in the Department of Computer Science and Engineering at the University of Notre Dame. His research lies in controllable knowledge-driven natural language processing, particularly in natural language generation. His research has been published in top-ranked NLP and data mining conferences such as ACL, EMNLP, AAAI, WWW, and CIKM. Additional information is available at https://wyu97.github.io/.

Meng Jiang is an assistant professor in the Department of Computer Science and Engineering at the University of Notre Dame. He received his B.E. and Ph.D. in Computer Science from Tsinghua University and was a postdoctoral research associate at the University of Illinois at Urbana-Champaign. His research interests focus on knowledge graph construction and natural language generation for news summarization and forum post generation. The awards he received include Notre Dame Faculty Award in 2019 and Best Paper Awards at ISDSA and KDD-DLG in 2020. Additional information is available at http://www.meng-jiang.com/.

Zhiting Hu is an assistant professor in Halicioğlu Data Science Institute at UC San Diego. He received his Ph.D. in Machine Learning from Carnegie Mellon University. His research interest lies in the broad area of natural language processing in particular controllable text generation, machine learning to enable training AI agents from all forms of experiences such as structured knowledge, ML systems and applications. His research was recognized with best demo nomination at ACL 2019 and outstanding paper award at ACL 2016. Additional information is available at http://www.cs.cmu.edu/?zhitingh/.

Qingyun Wang is a Ph.D. student in the Computer Science Department at the University of Illinois at Urbana-Champaign. His research lies in controllable knowledge-driven natural language generation, with a recent focus on the scientific paper generation. He served as a program committee in generation track for multiple conferences including ICML 2020, ACL 2019-2020, ICLR 2021, etc. He previously entered the finalist of the first Alexa Prize competition. Additional information is available at https://eaglew.github.io/.

Heng Ji is a professor at Computer Science Department of University of Illinois at Urbana-Champaign, and Amazon Scholar. She has published on Multimedia Multilingual Information Extraction and Knowledge-enriched NLG including technical paper generation, knowledge base description, and knowledge-aware image and video caption generation. The awards she received include "Young Scientist" by World Economic Fo- rum, "AIÕs 10 to Watch" Award by IEEE Intel- ligent Systems, NSF CAREER award, and ACL 2020 Best Demo Award. She has served as the Program Committee Co-Chair of many con- ferences including NAACL-HLT2018, and she is NAACL secretary 2020-2021. Additional information is available at https://blender.cs.illinois.edu/hengji.html.

Nazneen Rajani is a senior research scientist at Salesforce Research. She got her PhD in Computer Science from UT Austin in 2018. Several of her work has been published in ACL, EMNLP, NACCL, and IJCAI including work on generating explanations for commonsense and physical reasoning. Nazneen was one of the finalists for the VentureBeat Transform 2020 women in AI Research. Her work has been covered by several media outlets including Quanta Magazine, VentureBeat, SiliconAngle, ZDNet. More information on https://www.nazneenrajani.com.

3

Tutorials: Thursday, November 11

Overview

9:00 – 12:30 **Morning Tutorials**

Multi-Domain Multilingual Question Answering
Sebastian Ruder, Avirup Sil

Robustness and Adversarial Examples in Natural Language Processing
Kai-Wei Chang, He He, Robin Jia, Sameer Singh

Syntax in End-to-End Natural Language Processing
Hai Zhao, Rui Wang, Kehai Chen

Tutorial 4

Multi-Domain Multilingual Question Answering

Sebastian Ruder, **Avirup Sil**

Thursday, November 11, 2021, 9:00–12:30pm

Question answering (QA) is one of the most challenging and impactful tasks in natural language processing. Most research in QA and tutorials, however, has focused on the open-domain or monolingual setting while most real-world applications deal with specific domains or languages. In this tutorial, we attempt to bridge this gap. Firstly, we introduce standard benchmarks in multi-domain and multilingual QA. In both scenarios, we discuss state-of-the-art approaches that achieve impressive performance by either zero-shot learning or out-of-the-box training on open (and closed)-domain QA systems. Finally, we will present open research problems that this new research agenda poses such as multi-task learning, cross-lingual transfer learning, domain adaptation and training large scale pre-trained multilingual language models.

Sebastian Ruder is a research scientist at DeepMind where he works on transfer learning and multilingual natural language processing. He has been area chair in machine learning and multilinguality for major NLP conferences including ACL and EMNLP and has published papers on multilingual question answering (Artetxe et al., 2020; Hu et al., 2020). He was the Co-Program Chair for EurNLP 2019 and has co-organized the 4th Workshop on Representation Learning for NLP at ACL 2019. He has taught tutorials on "Transfer learning in natural language processing" and "Unsupervised Cross-lingual Representation Learning" at NAACL 2019 and ACL 2019 respectively. He has also co-organized and taught at the NLP Session at the Deep Learning Indaba 2018 and 2019.

Avirup Sil is a Research Scientist and the Team Lead for Question Answering in the Multilingual NLP group at IBM Research AI. His team (comprising of research scientists and engineers) works on research on industry scale NLP and Deep Learning algorithms. His teamÕs system called 'GAAMA' has obtained the top scores in public benchmark datasets (Kwiatkowski et al., 2019) and has published several papers on question answering (Chakravarti et al., 2019; Castelli et al., 2020; Glass et al., 2020). He is also the Chair of the NLP professional community of IBM. Avi is a Senior Program Committee Member and the Area Chair in Question Answering for major NLP conferences e.g. ACL, EMNLP, NAACL and has published several papers on Question Answering. He has taught a tutorial at ACL 2018 on "Entity Discovery and Linking". He has also organized the workshop on the "Relevance of Linguistic Structure in Neural NLP" at ACL 2018. He is also the track coordinator for the Entity Discovery and Linking track at the Text Analysis Conference.

Tutorial 5

Robustness and Adversarial Examples in Natural Language Processing

Kai-Wei Chang, **He He**, **Robin Jia**, **Sameer Singh**

Thursday, November 11, 2021, 9:00–12:30pm

Recent studies show that many NLP systems are sensitive and vulnerable to a small perturbation of inputs and do not generalize well across different datasets. This lack of robustness derails the use of NLP systems in real-world applications. This tutorial aims at bringing awareness of practical concerns about NLP robustness. It targets NLP researchers and practitioners who are interested in building reliable NLP systems. In particular, we will review recent studies on analyzing the weakness of NLP systems when facing adversarial inputs and data with a distribution shift. We will provide the audience with a holistic view of 1) how to use adversarial examples to examine the weakness of NLP models and facilitate debugging; 2) how to enhance the robustness of existing NLP models and defense against adversarial inputs; and 3) how the consideration of

---

Kai-Wei Chang is an assistant professor in the Department of Computer Science at the University of California Los Angeles. His research interests include designing robust, fair, and accountable machine learning methods for building reliable NLP systems (e.g., Alzantot et al., 2018; Shi et al., 2019). His awards include the EMNLP Best Long Paper Award (2017), the KDD Best Paper Award (2010), and the Sloan Research Fellowship (2021). Kai-Wei has given tutorials at NAACL 15, AAAI 16, FAccT18, EMNLP 19, AAAI 20, MLSS 21 on different research topics. Additional information is available at http://kwchang.net.

He He is an assistant professor in the Department of Computer Science and the Center for Data Science at the New York University. Her research interests include reliable natural language generation and robust learning algorithms that avoid spurious correlations in the data (e.g., He et al., 2019; Tu et al., 2020). She has given tutorials at NAACL 15 and EMNLP 19. Additional information is available at http://hhexiy.github.io.

Robin Jia is currently a visiting researcher at Facebook AI Research, and will be an assistant professor in the Department of Computer Science at the University of Southern California starting in the Autumn of 2021. His research focuses on making natural language processing models robust to unexpected test-time distribution shifts (e.g., Jia and Liang, 2017; Jia et al., 2019). RobinÕs work has received an Outstanding Paper Award at EMNLP 2017 and a Best Short Paper Award at ACL 2018. Additional information is available at https://robinjia.github.io.

Sameer Singh is an Assistant Professor of Computer Science at the University of California, Irvine. He is working on large-scale and interpretable machine learning models for NLP (e.g., Wallace et al., 2019a; Pezeshkpour et al., 2019). His work has received paper awards at ACL 2020, AKBC 2020, EMNLP 2019, ACL 2018, and KDD 2016. Sameer presented the Deep Adversarial Learning Tutorial (Wang et al., 2019) at NAACL 2019 and the Mining Knowledge Graphs from Text Tutorial at WSDM 2018 and AAAI 2017, along with tutorials on Interpretability and Explanations in upcoming NeurIPS 2020 and EMNLP 2020. Sameer has also received teaching awards at UCI. Website: http://sameersingh.org/.

robustness affects the real-world NLP applications used in our daily lives. We will conclude the tutorial by outlining future research directions in this area.

Tutorial 6

Syntax in End-to-End Natural Language Processing

Hai Zhao, **Rui Wang**, **Kehai Chen**

Thursday, November 11, 2021, 9:00–12:30pm

This tutorial surveys the latest technical progress of syntactic parsing and the role of syntax in end-to-end natural language processing (NLP) tasks, in which semantic role labeling (SRL) and machine translation (MT) are the representative NLP tasks that have always been beneficial from informative syntactic clues since a long time ago, though the advance from end-to-end deep learning models shows new results. In this tutorial, we will first introduce the background and the latest progress of syntactic parsing and SRL/NMT. Then, we will summarize the key evidence about the syntactic impacts over these two concerning tasks, and explore the behind reasons from both computational and linguistic background.

Hai Zhao is a professor at the Department of Computer Science and Engineering, Shanghai Jiao Tong University, China. His research interest is natural language processing. He has published more than 120 papers in ACL, EMNLP, COLING, ICLR, AAAI, IJCAI, and IEEE TKDE/TASLP. He won the first place in several NLP shared tasks, such as CoNLL and SIGHAN Bakeoff and top ranking in remarkable machine reading comprehension task leaderboards such as SQuAD2.0 and RACE. He has taught the course "natural language processing" in SJTU for more than 10 years. He is ACL-2017 area chair on parsing, and ACL- 2018/2019 (senior) area chairs on morphology and word segmentation.

Rui Wang is a tenured researcher at the Advanced Translation Technology Laboratory, National Institute of Information and Communications Technology (NICT), Japan. His research focuses on machine translation (MT), a classic task in NLP. His recent interests are traditional linguistic based and cutting-edge machine learning based approaches for MT. He (as the first or the corresponding authors) has published more than 30 MT papers in top-tier NLP/ML/AI conferences and journals, such as ACL, EMNLP, ICLR, AAAI, IJCAI, IEEE/ACM transactions, etc. He has also won several first places in top-tier MT shared tasks, such as WMT- 2018, WMT-2019, WMT-2020, etc. He has given several tutorial and invited talks in conferences, such as CWMT, CCL, etc. He served as the area chairs of ICLR-2021 and NAACL- 2021.

Kehai Chen is a postdoctoral researcher at the Advanced Translation Technology Laboratory, National Institute of Information and Communications Technology (NICT), Japan. His research focuses on linguistic-motivated machine translation (MT), a classic NLP task in AI. He has published more than 20 MT and NLP papers in top-tier NLP/ML/AI conferences and journals, such as ACL, ICLR, AAAI, EMNLP, IEEE/ACM Transactions on Audio, Speech, and Language Processing, ACM Transactions on Asian and Low-Resource Language Information Processing, etc. He served as a senior program committee of AAAI-2021.

4

Main Conference: Sunday, November 7

Overview

8:45 – 9:00 **Opening Remarks**

9:00 – 10:00 **Invited Talk 1 by Ido Dagan** *Bavaro 1, 2, 3*

10:00 – 10:30 **Coffee Break**

Session 1

10:30 – 12:00						
	Machine Translation and Multilinguality 1	Summarization 1	Information Extraction 1	Sentiment Analysis, Stylistic Analysis, and Argument Mining 1	Computational Social Science and Cultural Analytics	Efficient Methods for NLP 1
	Bavaro 1	*Bavaro 2*	*Bavaro 3*	*Punta Cana 1,2*	*Punta Cana 3,4*	*Santo Domingo*
	[Posters&Demos]: In-person Poster 1					*Bavaro 4*

Session 2

13:00 – 14:30						
	Machine Learning for NLP 1	Generation 1	Interpretability and Analysis of Models for NLP 1	NLP Applications 1	Linguistic Theories, Cognitive Modeling and Psycholinguistics	Information Retrieval and Text Mining
	Bavaro 1	*Bavaro 2*	*Bavaro 3*	*Punta Cana 1,2*	*Punta Cana 3,4*	*Santo Domingo*

14:30 – 14:45 **Mini Break**

Session 3

14:45 – 16:15						
	Machine Learning for NLP 2	Dialogue and Interactive Systems 1	Information Extraction 2	Resources and Evaluation 1	Discourse and Pragmatics	Semantics 1
	Bavaro 1	*Bavaro 2*	*Bavaro 3*	*Punta Cana 1,2*	*Punta Cana 3,4*	*Santo Domingo*
	[Posters&Demos]: In-person Poster 2					*Bavaro 4*

16:15 – 16:45 **Coffee Break**

Session 4

16:45 – 18:15						
	Machine Learning for NLP 3	Dialogue and Interactive Systems 2	Information Extraction 3	Ethics and NLP	Phonology, Morphology and Word Segmentation	Speech, Vision, Robotics, Multimodal Grounding 1
	Bavaro 1	*Bavaro 2*	*Bavaro 3*	*Punta Cana 1,2*	*Punta Cana 3,4*	*Santo Domingo*

19:00 – 21:00 **Virtual Poster Session I**

Keynote Address: Ido Dagan

Where next? Towards multi-text consumption via three inspired research lines

Sunday, November 7, 2021, 9:00–10:00am

Bavaro 1, 2, 3

Abstract: What effect does language have on people?

While the Dominican Republic is obviously my next exciting travel destination, in this talk I'll share what I consider as exciting destinations for next steps in NLP research. I'll start by pointing at a motivating grand application challenge: supporting effective human consumption of multi-text information – an invaluable goal which has seen very little progress since search engine inception. I'll then describe three individual, yet synergetic, research lines that were inspired by seeking this goal. First, supporting multi-text consumption is inherently an interactive process, where the user assists and directs the system in presenting most valuable information. As a first step, we propose a formulation of interactive summarization, turning it into a viable and measurable research task by extending summarization evaluation methods to the interactive setting. Second, presenting scattered information in a concise and consolidated manner requires extensive methods for linking cross-text information. Promoting such a research line, we propose several infrastructure contributions to cross-document coreference resolution, extend the scope of matching cross-text information to the interpretable levels of proposition spans and predicate-argument relations, and design a Cross-document Language Model (CDLM) which is geared for the multi-text setting. Lastly, we suggest that linking and consolidating multi-text information in a refined and controllable manner can benefit from some explicit interpretable representations of textual information. Rather than following traditional formal semantic representations, we propose a midway between those and opaque distributed neural representations. Text information is decomposed into a set of minimal natural language question-answer pairs, providing a generally appealing semi-structured representation for propositions in a single text, as well as a basis for aligning cross-text information units. Altogether, we advocate the promise of each individual research line for NLP progress, while suggesting human consumption of multi-text information as an inspiring research framework with a huge applied value.

Biography: Ido Dagan is a Professor at the Department of Computer Science at Bar-Ilan University, Israel, the founder of the Natural Language Processing (NLP) Lab at Bar-Ilan, the founder and head of the nationally funded Bar-Ilan University Data Science Institute, and a Fellow of the Association for Computational Linguistics (ACL). His interests are in applied semantic processing, focusing on textual inference, natural open semantic representations, consolidation and summarization of multi-text information, and interactive text summarization and exploration. Dagan and colleagues initiated and promoted textual entailment recognition (RTE, later aka NLI) as a generic empirical task. He was the President of the ACL in 2010 and served on its Executive Committee during 2008-2011. In that capacity, he led the establishment of the journal Transactions of the Association for Computational Linguistics, which became one of two premiere journals in NLP. Dagan received his B.A. summa cum laude and his Ph.D. (1992) in Computer Science from the Technion. He was a research fellow at the IBM Haifa Scientific Center (1991) and a Member of Technical Staff at ATT Bell Laboratories (1992-1994). During 1998-2003 he was co-founder and CTO of FocusEngine and VP of Technology of LingoMotors, and has been regularly consulting in the industry. His academic research has involved extensive industrial collaboration, including funds from IBM, Google, Thomson-Reuters, Bloomberg, Intel and Facebook, as well as collaboration with local companies under funded projects of the Israel Innovation Authority.

Session 1 Overview – Sunday, November 7, 2021

Track A	**Track B**	**Track C**	**Track D**	**Track E**	**Track F**	
Machine Translation and Multilinguality 1	*Summarization 1*	*Information Extraction 1*	*Sentiment Analysis, Stylistic Analysis, and Argument Mining 1*	*Computational Social Science and Cultural Analytics*	*Efficient Methods for NLP 1*	
AligNART: Non-autoregressive Neural Machine Translation by Jointly Learning to Estimate Alignment and Translate *Song, Kim, and Yoon*	Low-Resource Dialogue Summarization with Domain-Agnostic Multi-Source Pretraining *Zou et al.*	Variational Deep Logic Network for Joint Inference of Entities and Relations *Wang and Pan*	Beta Distribution Guided Aspect-aware Graph for Aspect Category Sentiment Analysis with Affective Knowledge *Liang et al.*	Idiosyncratic but not Arbitrary: Learning Idiolects in Online Registers Reveals Distinctive yet Consistent Individual Styles *Zhu and Jurgens*	Distilling Linguistic Context for Language Model Compression *Park, Kim, and Yang*	10:30
Zero-Shot Cross-Lingual Transfer of Neural Machine Translation with Multilingual Pretrained Encoders *Chen et al.*	Controllable Neural Dialogue Summarization with Personal Named Entity Planning *Liu and Chen*	Unsupervised Keyphrase Extraction by Jointly Modeling Local and Global Context *Liang et al.*	DILBERT: Customized Pre-Training for Domain Adaptation with Category Shift, with an Application to Aspect Extraction *Lekhtman, Ziser, and Reichart*	Narrative Theory for Computational Narrative Understanding *Piper, So, and Bamman*	Dynamic Knowledge Distillation for Pre-trained Language Models *Li et al.*	10:45
ERNIE-M: Enhanced Multilingual Representation by Aligning Cross-lingual Semantics with Monolingual Corpora *Ouyang et al.*	Fine-grained Factual Consistency Assessment for Abstractive Summarization Models *Zhang, Niu, and Wei*	Distantly Supervised Relation Extraction using Multi-Layer Revision Network and Confidence-based Multi-Instance Learning *Lin et al.*	Improving Multimodal fusion via Mutual Dependency Maximisation *Colombo et al.*	(Mis)alignment Between Stance Expressed in Social Media Data and Public Opinion Surveys *Joseph et al.*	Few-Shot Text Generation with Natural Language Instructions *Schick and Schütze*	11:00

	Track A	Track B	Track C	Track D	Track E	Track F
	Machine Translation and Multilinguality 1	*Summarization 1*	*Information Extraction 1*	*Sentiment Analysis, Stylistic Analysis, and Argument Mining 1*	*Computational Social Science and Cultural Analytics*	*Efficient Methods for NLP 1*
11:15	Cross Attention Augmented Transducer Networks for Simultaneous Translation *Liu et al.*	Decision-Focused Summarization *Hsu and Tan*	Logic-level Evidence Retrieval and Graph-based Verification Network for Table-based Fact Verification *Shi et al.*	Learning Implicit Sentiment in Aspect-based Sentiment Analysis with Supervised Contrastive Pre-Training *Li et al.*	How Does Counterfactually Augmented Data Impact Models for Social Computing Constructs? *Sen et al.*	SOM-NCSCM : An Efficient Neural Chinese Sentence Compression Model Enhanced with Self-Organizing Map *Zi et al.*
11:30	Translating Headers of Tabular Data: A Pilot Study of Schema Translation *Zhu et al.*	Multiplex Graph Neural Network for Extractive Text Summarization *Jing et al.*	A Partition Filter Network for Joint Entity and Relation Extraction *Yan et al.*	Progressive Self-Training with Discriminator for Aspect Term Extraction *Wang et al.*	Latent Hatred: A Benchmark for Understanding Implicit Hate Speech *ElSherief et al.*	Efficient Multi-Task Auxiliary Learning: Selecting Auxiliary Data by Feature Similarity *Kung et al.*
11:45	Towards Making the Most of Dialogue Characteristics for Neural Chat Translation *Liang et al.*	A Thorough Evaluation of Task-Specific Pretraining for Summarization *Rothe, Maynez, and Narayan*	TEBNER: Domain Specific Named Entity Recognition with Type Expanded Boundary-aware Network *Fang et al.*	Reinforced Counterfactual Data Augmentation for Dual Sentiment Classification *Chen, Xia, and Yu*	Aligning Faithful Interpretations with their Social Attribution *Jacovi and Goldberg*	Model Compression for Domain Adaptation through Causal Effect Estimation *Rotman, Feder, and Reichart*
11:50		HETFORMER: Heterogeneous Transformer with Sparse Attention for Long-Text Extractive Summarization *Liu et al.*				

Parallel Session 1

Session 1A: Machine Translation and Multilinguality 1

Bavaro 1 *Chair: Alexandra Birch*

AligNART: Non-autoregressive Neural Machine Translation by Jointly Learning to Estimate Alignment and Translate

Jongyoon Song, Sungwon Kim, and Sungroh Yoon 10:30–10:45

Non-autoregressive neural machine translation (NART) models suffer from the multi-modality problem which causes translation inconsistency such as token repetition. Most recent approaches have attempted to solve this problem by implicitly modeling dependencies between outputs. In this paper, we introduce AligNART, which leverages full alignment information to explicitly reduce the modality of the target distribution. AligNART divides the machine translation task into (i) alignment estimation and (ii) translation with aligned decoder inputs, guiding the decoder to focus on simplified one-to-one translation. To alleviate the alignment estimation problem, we further propose a novel alignment decomposition method. Our experiments show that AligNART outperforms previous non-iterative NART models that focus on explicit modality reduction on WMT14 EnDe and WMT16 Ro→En. Furthermore, AligNART achieves BLEU scores comparable to those of the state-of-the-art connectionist temporal classification based models on WMT14 EnDe. We also observe that AligNART effectively addresses the token repetition problem even without sequence-level knowledge distillation.

Zero-Shot Cross-Lingual Transfer of Neural Machine Translation with Multilingual Pretrained Encoders

Guanhua Chen et al. 10:45–11:00

Previous work mainly focuses on improving cross-lingual transfer for NLU tasks with a multilingual pretrained encoder (MPE), or improving the performance on supervised machine translation with BERT. However, it is under-explored that whether the MPE can help to facilitate the cross-lingual transferability of NMT model. In this paper, we focus on a zero-shot cross-lingual transfer task in NMT. In this task, the NMT model is trained with parallel dataset of only one language pair and an off-the-shelf MPE, then it is directly tested on zero-shot language pairs. We propose SixT, a simple yet effective model for this task. SixT leverages the MPE with a two-stage training schedule and gets further improvement with a position disentangled encoder and a capacity-enhanced decoder. Using this method, SixT significantly outperforms mBART, a pretrained multilingual encoder-decoder model explicitly designed for NMT, with an average improvement of 7.1 BLEU on zero-shot any-to-English test sets across 14 source languages. Furthermore, with much less training computation cost and training data, our model achieves better performance on 15 any-to-English test sets than CRISS and m2m-100, two strong multilingual NMT baselines.

ERNIE-M: Enhanced Multilingual Representation by Aligning Cross-lingual Semantics with Monolingual Corpora

Xuan Ouyang et al. 11:00–11:15

Recent studies have demonstrated that pre-trained cross-lingual models achieve impressive performance in downstream cross-lingual tasks. This improvement benefits from learning a large amount of monolingual and parallel corpora. Although it is generally acknowledged that parallel corpora are critical for improving the model performance, existing methods are often constrained by the size of parallel corpora, especially for low-resource languages. In this paper, we propose Ernie-M, a new training method that encourages the model to align the representation of multiple languages with monolingual corpora, to overcome the constraint that the parallel corpus size places on the model performance. Our key insight is to integrate back-translation into the pre-training process. We generate pseudo-parallel sentence pairs on a monolingual corpus to enable the learning of semantic alignments between different languages, thereby enhancing the semantic modeling of cross-lingual models. Experimental results show that Ernie-M outperforms existing cross-lingual models and delivers new state-of-the-art results in various cross-lingual downstream tasks. The codes and pre-trained models will be made publicly available.

Cross Attention Augmented Transducer Networks for Simultaneous Translation

Dan Liu et al. 11:15–11:30

This paper proposes a novel architecture, Cross Attention Augmented Transducer (CAAT), for simultaneous translation. The framework aims to jointly optimize the policy and translation models. To effectively consider all possible READ-WRITE simultaneous translation action paths, we adapt the online automatic speech recognition (ASR) model, RNN-T, but remove the strong monotonic constraint, which is critical for the translation task to consider reordering. To make CAAT work, we introduce a novel latency loss whose expectation can be optimized by a forward-backward algorithm. We implement CAAT with Transformer while the general CAAT architecture can also be implemented with other attention-based encoder-decoder frameworks. Experiments on both speech-to-text (S2T) and text-to-text (T2T) simultaneous translation tasks show that CAAT achieves significantly better latency-quality trade-offs compared to the state-of-the-art simultaneous translation approaches.

Translating Headers of Tabular Data: A Pilot Study of Schema Translation

Kunrui Zhu et al. 11:30–11:45

Schema translation is the task of automatically translating headers of tabular data from one language to another. High-quality schema translation plays an important role in cross-lingual table searching, understanding and analysis. Despite its importance, schema translation is not well studied in the community, and state-of-the-art neural machine translation models cannot work well on this task because of two intrinsic differences between plain text and tabular data: morphological difference and context difference. To facilitate the research study, we construct the first parallel dataset for schema translation, which consists of 3,158 tables with 11,979 headers written in 6 different languages, including English, Chinese, French, German, Spanish, and Japanese. Also, we propose the first schema translation model called CAST, which is a header-to-header neural machine translation model augmented with schema context. Specifically, we model a target header and its context as a directed graph to represent their entity types and relations. Then CAST encodes the graph with a relational-aware transformer and uses another transformer to decode the header in the target language. Experiments on our dataset demonstrate that CAST significantly outperforms state-of-the-art neural machine translation models. Our dataset will be released at https://github.com/microsoft/ContextualSP.

Towards Making the Most of Dialogue Characteristics for Neural Chat Translation
Yunlong Liang et al. 11:45–12:00

Neural Chat Translation (NCT) aims to translate conversational text between speakers of different languages. Despite the promising performance of sentence-level and context-aware neural machine translation models, there still remain limitations in current NCT models because the inherent dialogue characteristics of chat, such as dialogue coherence and speaker personality, are neglected. In this paper, we propose to promote the chat translation by introducing the modeling of dialogue characteristics into the NCT model. To this end, we design four auxiliary tasks including monolingual response generation, cross-lingual response generation, next utterance discrimination, and speaker identification. Together with the main chat translation task, we optimize the enhanced NCT model through the training objectives of all these tasks. By this means, the NCT model can be enhanced by capturing the inherent dialogue characteristics, thus generating more coherent and speaker-relevant translations. Comprehensive experiments on four language directions (English<->German and English<->Chinese) verify the effectiveness and superiority of the proposed approach.

Session 1B: Summarization 1

Bavaro 2 *Chair: Margot Mieskes*

Low-Resource Dialogue Summarization with Domain-Agnostic Multi-Source Pretraining

Yicheng Zou et al. 10:30–10:45

With the rapid increase in the volume of dialogue data from daily life, there is a growing demand for dialogue summarization. Unfortunately, training a large summarization model is generally infeasible due to the inadequacy of dialogue data with annotated summaries. Most existing works for low-resource dialogue summarization directly pretrain models in other domains, e.g., the news domain, but they generally neglect the huge difference between dialogues and conventional articles. To bridge the gap between out-of-domain pretraining and in-domain fine-tuning, in this work, we propose a multi-source pretraining paradigm to better leverage the external summary data. Specifically, we exploit large-scale in-domain non-summary data to separately pretrain the dialogue encoder and the summary decoder. The combined encoder-decoder model is then pretrained on the out-of-domain summary data using adversarial critics, aiming to facilitate domain-agnostic summarization. The experimental results on two public datasets show that with only limited training data, our approach achieves competitive performance and generalizes well in different dialogue scenarios.

Controllable Neural Dialogue Summarization with Personal Named Entity Planning

Zhengyuan Liu and Nancy Chen 10:45–11:00

In this paper, we propose a controllable neural generation framework that can flexibly guide dialogue summarization with personal named entity planning. The conditional sequences are modulated to decide what types of information or what perspective to focus on when forming summaries to tackle the under-constrained problem in summarization tasks. This framework supports two types of use cases: (1) Comprehensive Perspective, which is a general-purpose case with no user-preference specified, considering summary points from all conversational interlocutors and all mentioned persons; (2) Focus Perspective, positioning the summary based on a user-specified personal named entity, which could be one of the interlocutors or one of the persons mentioned in the conversation. During training, we exploit occurrence planning of personal named entities and coreference information to improve temporal coherence and to minimize hallucination in neural generation. Experimental results show that our proposed framework generates fluent and factually consistent summaries under various planning controls using both objective metrics and human evaluations.

Fine-grained Factual Consistency Assessment for Abstractive Summarization Models

Sen Zhang, Jianwei Niu, and Chuyuan Wei 11:00–11:15

Factual inconsistencies existed in the output of abstractive summarization models with original documents are frequently presented. Fact consistency assessment requires the reasoning capability to find subtle clues to identify whether a model-generated summary is consistent with the original document. This paper proposes a fine-grained two-stage Fact Consistency assessment framework for Summarization models (SumFC). Given a document and a summary sentence, in the first stage, SumFC selects the top-K most relevant sentences with the summary sentence from the document. In the second stage, the model performs fine-grained consistency reasoning at the sentence level, and then aggregates all sentences' consistency scores to obtain the final assessment result. We get the training data pairs by data synthesis and adopt contrastive loss of data pairs to help the model identify subtle cues. Experiment results show that SumFC has made a significant improvement over the previous state-of-the-art methods. Our experiments also indicate that SumFC distinguishes detailed differences better.

Decision-Focused Summarization

Chao-Chun Hsu and Chenhao Tan 11:15–11:30

Relevance in summarization is typically de- fined based on textual information alone, without incorporating insights about a particular decision. As a result, to support risk analysis of pancreatic cancer, summaries of medical notes may include irrelevant information such as a knee injury. We propose a novel problem, decision-focused summarization, where the goal is to summarize relevant information for a decision. We leverage a predictive model that makes the decision based on the full text to provide valuable insights on how a decision can be inferred from text. To build a summary, we then select representative sentences that lead to similar model decisions as using the full text while accounting for textual non-redundancy. To evaluate our method (DecSum), we build a testbed where the task is to summarize the first ten reviews of a restaurant in support of predicting its future rating on Yelp. DecSum substantially outperforms text-only summarization methods and model-based explanation methods in decision faithfulness and representativeness. We further demonstrate that DecSum is the only method that enables humans to outperform random chance in predicting which restaurant will be better rated in the future.

Multiplex Graph Neural Network for Extractive Text Summarization

Baoyu Jing et al. 11:30–11:40

Extractive text summarization aims at extracting the most representative sentences from a given document as its summary. To extract a good summary from a long text document, sentence embedding plays an important role. Recent studies have leveraged graph neural networks to capture the inter-sentential relationship (e.g., the discourse graph) within the documents to learn contextual sentence embedding. However, those approaches neither consider multiple types of inter-sentential relationships (e.g., semantic similarity and natural connection relationships), nor model intra-sentential relationships (e.g, semantic similarity and syntactic

relationship among words). To address these problems, we propose a novel Multiplex Graph Convolutional Network (Multi-GCN) to jointly model different types of relationships among sentences and words. Based on Multi-GCN, we propose a Multiplex Graph Summarization (Multi-GraS) model for extractive text summarization. Finally, we evaluate the proposed models on the CNN/DailyMail benchmark dataset to demonstrate effectiveness of our method.

A Thorough Evaluation of Task-Specific Pretraining for Summarization
Sascha Rothe, Joshua Maynez, and Shashi Narayan 11:40–11:50

Task-agnostic pretraining objectives like masked language models or corrupted span prediction are applicable to a wide range of NLP downstream tasks (Raffel et al.,2019), but are outperformed by task-specific pretraining objectives like predicting extracted gap sentences on summarization (Zhang et al.,2020). We compare three summarization specific pretraining objectives with the task agnostic corrupted span prediction pretraining in controlled study. We also extend our study to a low resource and zero shot setup, to understand how many training examples are needed in order to ablate the task-specific pretraining without quality loss. Our results show that task-agnostic pretraining is sufficient for most cases which hopefully reduces the need for costly task-specific pretraining. We also report new state-of-the-art number for two summarization task using a T5 model with 11 billion parameters and an optimal beam search length penalty.

HETFORMER: Heterogeneous Transformer with Sparse Attention for Long-Text Extractive Summarization
Ye Liu et al. 11:50–12:00

To capture the semantic graph structure from raw text, most existing summarization approaches are built on GNNs with a pre-trained model. However, these methods suffer from cumbersome procedures and inefficient computations for long-text documents. To mitigate these issues, this paper proposes HetFormer, a Transformer-based pre-trained model with multi-granularity sparse attentions for long-text extractive summarization. Specifically, we model different types of semantic nodes in raw text as a potential heterogeneous graph and directly learn heterogeneous relationships (edges) among nodes by Transformer. Extensive experiments on both single- and multi-document summarization tasks show that HetFormer achieves state-of-the-art performance in Rouge F1 while using less memory and fewer parameters.

Session 1C: Information Extraction 1

Bavaro 3 *Chair: Jie Yang*

Variational Deep Logic Network for Joint Inference of Entities and Relations

Wenya Wang and Sinno Jialin Pan 10:30–10:45

Currently, deep learning models have been widely adopted and achieved promising results on various application domains. Despite their intriguing performance, most deep learning models function as black boxes, lacking explicit reasoning capabilities and explanations, which are usually essential for complex problems. Take joint inference in information extraction as an example. This task requires the identification of multiple structured knowledge from texts, which is inter-correlated, including entities, events, and the relationships between them. Various deep neural networks have been proposed to jointly perform entity extraction and relation prediction, which only propagate information implicitly via representation learning. However, they fail to encode the intensive correlations between entity types and relations to enforce their coexistence. On the other hand, some approaches adopt rules to explicitly constrain certain relational facts, although the separation of rules with representation learning usually restrains the approaches with error propagation. Moreover, the predefined rules are inflexible and might result in negative effects when data is noisy. To address these limitations, we propose a variational deep logic network that incorporates both representation learning and relational reasoning via the variational EM algorithm. The model consists of a deep neural network to learn high-level features with implicit interactions via the self-attention mechanism and a relational logic network to explicitly exploit target interactions. These two components are trained interactively to bring the best of both worlds. We conduct extensive experiments ranging from fine-grained sentiment terms extraction, end-to-end relation prediction, to end-to-end event extraction to demonstrate the effectiveness of our proposed method.

Unsupervised Keyphrase Extraction by Jointly Modeling Local and Global Context

Xinnian Liang et al. 10:45–11:00

Embedding based methods are widely used for unsupervised keyphrase extraction (UKE) tasks. Generally, these methods simply calculate similarities between phrase embeddings and document embedding, which is insufficient to capture different context for a more effective UKE model. In this paper, we propose a novel method for UKE, where local and global contexts are jointly modeled. From a global view, we calculate the similarity between a certain phrase and the whole document in the vector space as transitional embedding based models do. In terms of the local view, we first build a graph structure based on the document where phrases are regarded as vertices and the edges are similarities between vertices. Then, we proposed a new centrality computation method to capture local salient information based on the graph structure. Finally, we further combine the modeling of global and local context for ranking. We evaluate our models on three public benchmarks (Inspec, DUC 2001, SemEval 2010) and compare with existing state-of-the-art models. The results show that our model outperforms most models while generalizing better on input documents with different domains and length. Additional ablation study shows that both the local and global information is crucial for unsupervised keyphrase extraction tasks.

Distantly Supervised Relation Extraction using Multi-Layer Revision Network and Confidence-based Multi-Instance Learning

Xiangyu Lin et al. 11:00–11:15

Distantly supervised relation extraction is widely used in the construction of knowledge bases due to its high efficiency. However, the automatically obtained instances are of low quality with numerous irrelevant words. In addition, the strong assumption of distant supervision leads to the existence of noisy sentences in the sentence bags. In this paper, we propose a novel Multi-Layer Revision Network (MLRN) which alleviates the effects of word-level noise by emphasizing inner-sentence correlations before extracting relevant information within sentences. Then, we devise a balanced and noise-resistant Confidence-based Multi-Instance Learning (CMIL) method to filter out noisy sentences as well as assign proper weights to relevant ones. Extensive experiments on two New York Times (NYT) datasets demonstrate that our approach achieves significant improvements over the baselines.

Logic-level Evidence Retrieval and Graph-based Verification Network for Table-based Fact Verification

Qi Shi et al. 11:15–11:30

Table-based fact verification task aims to verify whether the given statement is supported by the given semi-structured table. Symbolic reasoning with logical operations plays a crucial role in this task. Existing methods leverage programs that contain rich logical information to enhance the verification process. However, due to the lack of fully supervised signals in the program generation process, spurious programs can be derived and employed, which leads to the inability of the model to catch helpful logical operations. To address the aforementioned problems, in this work, we formulate the table-based fact verification task as an evidence retrieval and reasoning framework, proposing the Logic-level Evidence Retrieval and Graph-based Verification network (LERGV). Specifically, we first retrieve logic-level program-like evidence from the given table and statement as supplementary evidence for the table. After that, we construct a logic-level graph to capture the logical relations between entities and functions in the retrieved evidence, and design a graph-based verification network to perform logic-level graph-based reasoning based on the constructed graph to classify the final entailment relation. Experimental results on the large-scale benchmark TABFACT show the effectiveness of the proposed approach.

A Partition Filter Network for Joint Entity and Relation Extraction
Zhiheng Yan et al. 11:30–11:45

In joint entity and relation extraction, existing work either sequentially encode task-specific features, leading to an imbalance in inter-task feature interaction where features extracted later have no direct contact with those that come first. Or they encode entity features and relation features in a parallel manner, meaning that feature representation learning for each task is largely independent of each other except for input sharing. We propose a partition filter network to model two-way interaction between tasks properly, where feature encoding is decomposed into two steps: partition and filter. In our encoder, we leverage two gates: entity and relation gate, to segment neurons into two task partitions and one shared partition. The shared partition represents inter-task information valuable to both tasks and is evenly shared across two tasks to ensure proper two-way interaction. The task partitions represent intra-task information and are formed through concerted efforts of both gates, making sure that encoding of task-specific features is dependent upon each other. Experiment results on six public datasets show that our model performs significantly better than previous approaches. In addition, contrary to what previous work has claimed, our auxiliary experiments suggest that relation prediction is contributory to named entity prediction in a non-negligible way. The source code can be found at https://github.com/Coopercoppers/PFN.

TEBNER: Domain Specific Named Entity Recognition with Type Expanded Boundary-aware Network
Zheng Fang et al. 11:45–12:00

To alleviate label scarcity in Named Entity Recognition (NER) task, distantly supervised NER methods are widely applied to automatically label data and identify entities. Although the human effort is reduced, the generated incomplete and noisy annotations pose new challenges for learning effective neural models. In this paper, we propose a novel dictionary extension method which extracts new entities through the type expanded model. Moreover, we design a multi-granularity boundary-aware network which detects entity boundaries from both local and global perspectives. We conduct experiments on different types of datasets, the results show that our model outperforms previous state-of-the-art distantly supervised systems and even surpasses the supervised models.

Session 1D: Sentiment Analysis, Stylistic Analysis, and Argument Mining 1

Punta Cana 1,2 *Chair: Ivan Habernal*

Beta Distribution Guided Aspect-aware Graph for Aspect Category Sentiment Analysis with Affective Knowledge

Bin Liang et al. 10:30–10:45

In this paper, we investigate the Aspect Category Sentiment Analysis (ACSA) task from a novel perspective by exploring a Beta Distribution guided aspect-aware graph construction based on external knowledge. That is, we are no longer entangled about how to laboriously search the sentiment clues for coarse-grained aspects from the context, but how to preferably find the words highly related to the aspects in the context and determine their importance based on the public knowledge base. In this way, the contextual sentiment clues can be explicitly tracked in ACSA for the aspects in the light of these aspect-related words. To be specific, we first regard each aspect as a pivot to derive aspect-aware words that are highly related to the aspect from external affective commonsense knowledge. Then, we employ Beta Distribution to educe the aspect-aware weight, which reflects the importance to the aspect, for each aspect-aware word. Afterward, the aspect-aware words are served as the substitutes of the coarse-grained aspect to construct graphs for leveraging the aspect-related contextual sentiment dependencies in ACSA. Experiments on 6 benchmark datasets show that our approach significantly outperforms the state-of-the-art baseline methods.

DILBERT: Customized Pre-Training for Domain Adaptation with Category Shift, with an Application to Aspect Extraction

Entony Lekhtman, Yftah Ziser, and Roi Reichart 10:45–11:00

The rise of pre-trained language models has yielded substantial progress in the vast majority of Natural Language Processing (NLP) tasks. However, a generic approach towards the pre-training procedure can naturally be sub-optimal in some cases. Particularly, fine-tuning a pre-trained language model on a source domain and then applying it to a different target domain, results in a sharp performance decline of the eventual classifier for many source-target domain pairs. Moreover, in some NLP tasks, the output categories substantially differ between domains, making adaptation even more challenging. This, for example, happens in the task of aspect extraction, where the aspects of interest of reviews of, e.g., restaurants or electronic devices may be very different. This paper presents a new fine-tuning scheme for BERT, which aims to address the above challenges. We name this scheme DILBERT: Domain Invariant Learning with BERT, and customize it for aspect extraction in the unsupervised domain adaptation setting. DILBERT harnesses the categorical information of both the source and the target domains to guide the pre-training process towards a more domain and category invariant representation, thus closing the gap between the domains. We show that DILBERT yields substantial improvements over state-of-the-art baselines while using a fraction of the unlabeled data, particularly in more challenging domain adaptation setups.

Improving Multimodal fusion via Mutual Dependency Maximisation

Pierre Colombo et al. 11:00–11:15

Multimodal sentiment analysis is a trending area of research, and multimodal fusion is one of its most active topic. Acknowledging humans communicate through a variety of channels (i.e visual, acoustic, linguistic), multimodal systems aim at integrating different unimodal representations into a synthetic one. So far, a consequent effort has been made on developing complex architectures allowing the fusion of these modalities. However, such systems are mainly trained by minimising simple losses such as L_1 or cross-entropy. In this work, we investigate unexplored penalties and propose a set of new objectives that measure the dependency between modalities. We demonstrate that our new penalties lead to a consistent improvement (up to 4.3 on accuracy) across a large variety of state-of-the-art models on two well-known sentiment analysis datasets: `CMU-MOSI` and `CMU-MOSEI`. Our method not only achieves a new SOTA on both datasets but also produces representations that are more robust to modality drops. Finally, a by-product of our methods includes a statistical network which can be used to interpret the high dimensional representations learnt by the model.

Learning Implicit Sentiment in Aspect-based Sentiment Analysis with Supervised Contrastive Pre-Training

Zhengyan Li et al. 11:15–11:30

Aspect-based sentiment analysis aims to identify the sentiment polarity of a specific aspect in product reviews. We notice that about 30% of reviews do not contain obvious opinion words, but still convey clear human-aware sentiment orientation, which is known as implicit sentiment. However, recent neural network-based approaches paid little attention to implicit sentiment entailed in the reviews. To overcome this issue, we adopt Supervised Contrastive Pre-training on large-scale sentiment-annotated corpora retrieved from in-domain language resources. By aligning the representation of implicit sentiment expressions to those with the same sentiment label, the pre-training process leads to better capture of both implicit and explicit sentiment orientation towards aspects in reviews. Experimental results show that our method achieves state-of-the-art performance on SemEval2014 benchmarks, and comprehensive analysis validates its effectiveness on learning implicit sentiment.

Progressive Self-Training with Discriminator for Aspect Term Extraction

Qianlong Wang et al. 11:30–11:45

Aspect term extraction aims to extract aspect terms from a review sentence that users have expressed opinions on. One of the remaining challenges for aspect term extraction resides in the lack of sufficient annotated data. While self-training is potentially an effective method to address this issue, the pseudo-labels it yields on unlabeled data could induce noise. In this paper, we use two means to alleviate the noise in the pseudo-labels. One is that inspired by the curriculum learning, we refine the conventional self-training to progressive self-training. Specifically, the base model infers pseudo-labels on a progressive subset at each iteration, where samples in the subset become harder and more numerous as the iteration proceeds. The other is that we use a discriminator to filter the noisy pseudo-labels. Experimental results on four SemEval datasets show that our model significantly outperforms the previous baselines and achieves state-of-the-art performance.

Reinforced Counterfactual Data Augmentation for Dual Sentiment Classification
Hao Chen, Rui Xia, and Jianfei Yu 11:45–12:00

Data augmentation and adversarial perturbation approaches have recently achieved promising results in solving the over-fitting problem in many natural language processing (NLP) tasks including sentiment classification. However, existing studies aimed to improve the generalization ability by augmenting the training data with synonymous examples or adding random noises to word embeddings, which cannot address the spurious association problem. In this work, we propose an end-to-end reinforcement learning framework, which jointly performs counterfactual data generation and dual sentiment classification. Our approach has three characteristics:1) the generator automatically generates massive and diverse antonymous sentences; 2) the discriminator contains a original-side sentiment predictor and an antonymous-side sentiment predictor, which jointly evaluate the quality of the generated sample and help the generator iteratively generate higher-quality antonymous samples; 3) the discriminator is directly used as the final sentiment classifier without the need to build an extra one. Extensive experiments show that our approach outperforms strong data augmentation baselines on several benchmark sentiment classification datasets. Further analysis confirms our approach's advantages in generating more diverse training samples and solving the spurious association problem in sentiment classification.

Session 1E: Computational Social Science and Cultural Analytics

Punta Cana 3,4 *Chair: David Jurgens*

Idiosyncratic but not Arbitrary: Learning Idiolects in Online Registers Reveals Distinctive yet Consistent Individual Styles

Jian Zhu and David Jurgens 10:30–10:45

An individual's variation in writing style is often a function of both social and personal attributes. While structured social variation has been extensively studied, e.g., gender based variation, far less is known about how to characterize individual styles due to their idiosyncratic nature. We introduce a new approach to studying idiolects through a massive cross-author comparison to identify and encode stylistic features. The neural model achieves strong performance at authorship identification on short texts and through an analogy-based probing task, showing that the learned representations exhibit surprising regularities that encode qualitative and quantitative shifts of idiolectal styles. Through text perturbation, we quantify the relative contributions of different linguistic elements to idiolectal variation. Furthermore, we provide a description of idiolects through measuring inter- and intra-author variation, showing that variation in idiolects is often distinctive yet consistent.

Narrative Theory for Computational Narrative Understanding

Andrew Piper, Richard Jean So, and David Bamman 10:45–11:00

Over the past decade, the field of natural language processing has developed a wide array of computational methods for reasoning about narrative, including summarization, commonsense inference, and event detection. While this work has brought an important empirical lens for examining narrative, it is by and large divorced from the large body of theoretical work on narrative within the humanities, social and cognitive sciences. In this position paper, we introduce the dominant theoretical frameworks to the NLP community, situate current research in NLP within distinct narratological traditions, and argue that linking computational work in NLP to theory opens up a range of new empirical questions that would both help advance our understanding of narrative and open up new practical applications.

(Mis)alignment Between Stance Expressed in Social Media Data and Public Opinion Surveys

Kenneth Joseph et al. 11:00–11:15

Stance detection, which aims to determine whether an individual is for or against a target concept, promises to uncover public opinion from large streams of social media data. Yet even human annotation of social media content does not always capture "stance" as measured by public opinion polls. We demonstrate this by directly comparing an individual's self-reported stance to the stance inferred from their social media data. Leveraging a longitudinal public opinion survey with respondent Twitter handles, we conducted this comparison for 1,129 individuals across four salient targets. We find that recall is high for both "Pro" and "Anti" stance classifications but precision is variable in a number of cases. We identify three factors leading to the disconnect between text and author stance: temporal inconsistencies, differences in constructs, and measurement errors from both survey respondents and annotators. By presenting a framework for assessing the limitations of stance detection models, this work provides important insight into what stance detection truly measures.

How Does Counterfactually Augmented Data Impact Models for Social Computing Constructs?

Indira Sen et al. 11:15–11:30

As NLP models are increasingly deployed in socially situated settings such as online abusive content detection, it is crucial to ensure that these models are robust. One way of improving model robustness is to generate counterfactually augmented data (CAD) for training models that can better learn to distinguish between core features and data artifacts. While models trained on this type of data have shown promising out-of-domain generalizability, it is still unclear what the sources of such improvements are. We investigate the benefits of CAD for social NLP models by focusing on three social computing constructs — sentiment, sexism, and hate speech. Assessing the performance of models trained with and without CAD across different types of datasets, we find that while models trained on CAD show lower in-domain performance, they generalize better out-of-domain. We unpack this apparent discrepancy using machine explanations and find that CAD reduces model reliance on spurious features. Leveraging a novel typology of CAD to analyze their relationship with model performance, we find that CAD which acts on the construct directly or a diverse set of CAD leads to higher performance.

Latent Hatred: A Benchmark for Understanding Implicit Hate Speech

Mai ElSherief et al. 11:30–11:45

Hate speech has grown significantly on social media, causing serious consequences for victims of all demographics. Despite much attention being paid to characterize and detect discriminatory speech, most work has focused on explicit or overt hate speech, failing to address a more pervasive form based on coded or indirect language. To fill this gap, this work introduces a theoretically-justified taxonomy of implicit hate speech and a benchmark corpus with fine-grained labels for each message and its implication. We present systematic analyses of our dataset using contemporary baselines to detect and explain implicit hate speech, and we discuss key features that challenge existing models. This dataset will continue to serve as a useful benchmark for understanding this multifaceted issue.

Aligning Faithful Interpretations with their Social Attribution
Alon Jacovi and Yoav Goldberg 11:45–12:00

We find that the requirement of model interpretations to be faithful is vague and incomplete. With interpretation by textual highlights as a case-study, we present several failure cases. Borrowing concepts from social science, we identify that the problem is a misalignment between the causal chain of decisions (causal attribution) and the attribution of human behavior to the interpretation (social attribution). We re-formulate faithfulness as an accurate attribution of causality to the model, and introduce the concept of aligned faithfulness: faithful causal chains that are aligned with their expected social behavior. The two steps of causal attribution and social attribution together complete the process of explaining behavior. With this formalization, we characterize various failures of misaligned faithful highlight interpretations, and propose an alternative causal chain to remedy the issues. Finally, we implement highlight explanations of the proposed causal format using contrastive explanations.

Session 1F: Efficient Methods for NLP 1

Santo Domingo *Chair: Tong Xiao*

Distilling Linguistic Context for Language Model Compression

Geondo Park, Gyeongman Kim, and Eunho Yang 10:30–10:45

A computationally expensive and memory intensive neural network lies behind the recent success of language representation learning. Knowledge distillation, a major technique for deploying such a vast language model in resource-scarce environments, transfers the knowledge on individual word representations learned without restrictions. In this paper, inspired by the recent observations that language representations are relatively positioned and have more semantic knowledge as a whole, we present a new knowledge distillation objective for language representation learning that transfers the contextual knowledge via two types of relationships across representations: Word Relation and Layer Transforming Relation. Unlike other recent distillation techniques for the language models, our contextual distillation does not have any restrictions on architectural changes between teacher and student. We validate the effectiveness of our method on challenging benchmarks of language understanding tasks, not only in architectures of various sizes but also in combination with DynaBERT, the recently proposed adaptive size pruning method.

Dynamic Knowledge Distillation for Pre-trained Language Models

Lei Li et al. 10:45–11:00

Knowledge distillation~(KD) has been proved effective for compressing large-scale pre-trained language models. However, existing methods conduct KD statically, e.g., the student model aligns its output distribution to that of a selected teacher model on the pre-defined training dataset. In this paper, we explore whether a dynamic knowledge distillation that empowers the student to adjust the learning procedure according to its competency, regarding the student performance and learning efficiency. We explore the dynamical adjustments on three aspects: teacher model adoption, data selection, and KD objective adaptation. Experimental results show that (1) proper selection of teacher model can boost the performance of student model; (2) conducting KD with 10% informative instances achieves comparable performance while greatly accelerates the training; (3) the student performance can be boosted by adjusting the supervision contribution of different alignment objective. We find dynamic knowledge distillation is promising and provide discussions on potential future directions towards more efficient KD methods.

Few-Shot Text Generation with Natural Language Instructions

Timo Schick and Hinrich Schütze 11:00–11:15

Providing pretrained language models with simple task descriptions in natural language enables them to solve some tasks in a fully unsupervised fashion. Moreover, when combined with regular learning from examples, this idea yields impressive few-shot results for a wide range of text classification tasks. It is also a promising direction to improve data efficiency in generative settings, but there are several challenges to using a combination of task descriptions and example-based learning for text generation. In particular, it is crucial to find task descriptions that are easy to understand for the pretrained model and to ensure that it actually makes good use of them; furthermore, effective measures against overfitting have to be implemented. In this paper, we show how these challenges can be tackled: We introduce GenPET, a method for text generation that is based on pattern-exploiting training, a recent approach for combining textual instructions with supervised learning that only works for classification tasks. On several summarization and headline generation datasets, GenPET gives consistent improvements over strong baselines in few-shot settings.

SOM-NCSCM : An Efficient Neural Chinese Sentence Compression Model Enhanced with Self-Organizing Map

Kangli Zi et al. 11:15–11:30

Sentence Compression (SC), which aims to shorten sentences while retaining important words that express the essential meanings, has been studied for many years in many languages, especially in English. However, improvements on Chinese SC task are still quite few due to several difficulties: scarce of parallel corpora, different segmentation granularity of Chinese sentences, and imperfect performance of syntactic analyses. Furthermore, entire neural Chinese SC models have been under-investigated so far. In this work, we construct an SC dataset of Chinese colloquial sentences from a real-life question answering system in the telecommunication domain, and then, we propose a neural Chinese SC model enhanced with a Self-Organizing Map (SOM-NCSCM), to gain a valuable insight from the data and improve the performance of the whole neural Chinese SC model in a valid manner. Experimental results show that our SOM-NCSCM can significantly benefit from the deep investigation of similarity among data, and achieve a promising F1 score of 89.655 and BLEU4 score of 70.116, which also provides a baseline for further research on Chinese SC task.

Efficient Multi-Task Auxiliary Learning: Selecting Auxiliary Data by Feature Similarity

Po-Nien Kung et al. 11:30–11:45

Multi-task auxiliary learning utilizes a set of relevant auxiliary tasks to improve the performance of a primary task. A common usage is to manually select multiple auxiliary tasks for multi-task learning on all data, which raises two issues: (1) selecting beneficial auxiliary tasks for a primary task is nontrivial; (2) when the auxiliary datasets are large, training on all data becomes time-expensive and impractical. Therefore, this paper focuses on addressing these problems and proposes a time-efficient sampling method to select the data that is most relevant to the primary task. The proposed method allows us to only train on the most beneficial sub-datasets

from the auxiliary tasks, achieving efficient multi-task auxiliary learning. The experiments on three benchmark datasets (RTE, MRPC, STS-B) show that our method significantly outperforms random sampling and ST-DNN. Also, by applying our method, the model can surpass fully-trained MT-DNN on RTE, MRPC, STS-B, using only 50%, 66%, and 1% of data, respectively.

Model Compression for Domain Adaptation through Causal Effect Estimation
Guy Rotman, Amir Feder, and Roi Reichart 11:45–12:00

Recent improvements in the predictive quality of natural language processing systems are often dependent on a substantial increase in the number of model parameters. This has led to various attempts of compressing such models, but existing methods have not considered the differences in the predictive power of various model components or in the generalizability of the compressed models. To understand the connection between model compression and out-of-distribution generalization, we define the task of compressing language representation models such that they perform best in a domain adaptation setting. We choose to address this problem from a causal perspective, attempting to estimate the average treatment effect (ATE) of a model component, such as a single layer, on the model's predictions. Our proposed ATE-guided Model Compression scheme (AMoC), generates many model candidates, differing by the model components that were removed. Then, we select the best candidate through a stepwise regression model that utilizes the ATE to predict the expected performance on the target domain. AMoC outperforms strong baselines on dozens of domain pairs across three text classification and sequence tagging tasks.

Poster and Demo session 1

In-person Poster Session 1

Bavaro 4 *10:30–12:00*

GOLD: Improving Out-of-Scope Detection in Dialogues using Data Augmentation

Derek Chen and Zhou Yu

Practical dialogue systems require robust methods of detecting out-of-scope (OOS) utterances to avoid conversational breakdowns and related failure modes. Directly training a model with labeled OOS examples yields reasonable performance, but obtaining such data is a resource-intensive process. To tackle this limited-data problem, previous methods focus on better modeling the distribution of in-scope (INS) examples. We introduce GOLD as an orthogonal technique that augments existing data to train better OOS detectors operating in low-data regimes. GOLD generates pseudo-labeled candidates using samples from an auxiliary dataset and keeps only the most beneficial candidates for training through a novel filtering mechanism. In experiments across three target benchmarks, the top GOLD model outperforms all existing methods on all key metrics, achieving relative gains of 52.4%, 48.9% and 50.3% against median baseline performance. We also analyze the unique properties of OOS data to identify key factors for optimally applying our proposed method.

Graph Based Network with Contextualized Representations of Turns in Dialogue

Bongseok Lee and Yong Suk Choi

Dialogue-based relation extraction (RE) aims to extract relation(s) between two arguments that appear in a dialogue. Because dialogues have the characteristics of high personal pronoun occurrences and low information density, and since most relational facts in dialogues are not supported by any single sentence, dialogue-based relation extraction requires a comprehensive understanding of dialogue. In this paper, we propose the TUrn COntext awaRE Graph Convolutional Network (TUCORE-GCN) modeled by paying attention to the way people understand dialogues. In addition, we propose a novel approach which treats the task of emotion recognition in conversations (ERC) as a dialogue-based RE. Experiments on a dialogue-based RE dataset and three ERC datasets demonstrate that our model is very effective in various dialogue-based natural language understanding tasks. In these experiments, TUCORE-GCN outperforms the state-of-the-art models on most of the benchmark datasets. Our code is available at https://github.com/BlackNoodle/TUCORE-GCN.

Automatically Exposing Problems with Neural Dialog Models

Dian Yu and Kenji Sagae

Neural dialog models are known to suffer from problems such as generating unsafe and inconsistent responses. Even though these problems are crucial and prevalent, they are mostly manually identified by model designers through interactions. Recently, some research instructs crowdworkers to goad the bots into triggering such problems. However, humans leverage superficial clues such as hate speech, while leaving systematic problems undercover. In this paper, we propose two methods including reinforcement learning to automatically trigger a dialog model into generating problematic responses. We show the effect of our methods in exposing safety and contradiction issues with state-of-the-art dialog models.

Event Coreference Data (Almost) for Free: Mining Hyperlinks from Online News

Michael Bugert and Iryna Gurevych

Cross-document event coreference resolution (CDCR) is the task of identifying which event mentions refer to the same events throughout a collection of documents. Annotating CDCR data is an arduous and expensive process, explaining why existing corpora are small and lack domain coverage. To overcome this bottleneck, we automatically extract event coreference data from hyperlinks in online news: When referring to a significant real-world event, writers often add a hyperlink to another article covering this event. We demonstrate that collecting hyperlinks which point to the same article(s) produces extensive and high-quality CDCR data and create a corpus of 2M documents and 2.7M silver-standard event mentions called HyperCoref. We evaluate a state-of-the-art system on three CDCR corpora and find that models trained on small subsets of HyperCoref are highly competitive, with performance similar to models trained on gold-standard data. With our work, we free CDCR research from depending on costly human-annotated training data and open up possibilities for research beyond English CDCR, as our data extraction approach can be easily adapted to other languages.

Inducing Stereotypical Character Roles from Plot Structure

Labiba Jahan, Rahul Mittal, and Mark Finlayson

Stereotypical character roles-also known as archetypes or dramatis personae-play an important function in narratives: they facilitate efficient communication with bundles of default characteristics and associations and ease understanding of those characters' roles in the overall narrative. We present a fully unsupervised k-means clustering approach for learning stereotypical roles given only structural plot information. We demonstrate the technique on Vladimir Propp's structural theory of Russian folktales (captured in the extended ProppLearner corpus, with 46 tales), showing that our approach can induce six out of seven of Propp's dramatis personae with F1 measures of up to 0.70 (0.58 average), with an additional category for minor characters. We have explored various feature sets and variations of a cluster evaluation method. The best-performing feature set comprises plot functions, unigrams, tf-idf weights, and embeddings over coreference chain heads. Roles that are mentioned more often (Hero, Villain), or have clearly distinct plot patterns (Princess) are more strongly

differentiated than less frequent or distinct roles (Dispatcher, Helper, Donor). Detailed error analysis suggests that the quality of the coreference chain and plot functions annotations are critical for this task. We provide all our data and code for reproducibility.

Multitask Semi-Supervised Learning for Class-Imbalanced Discourse Classification
Alexander Spangher et al.

As labeling schemas evolve over time, small differences can render datasets following older schemas unusable. This prevents researchers from building on top of previous annotation work and results in the existence, in discourse learning in particular, of many small class-imbalanced datasets. In this work, we show that a multitask learning approach can combine discourse datasets from similar and diverse domains to improve discourse classification. We show an improvement of 4.9% Micro F1-score over current state-of-the-art benchmarks on the *NewsDiscourse* dataset, one of the largest discourse datasets recently published, due in part to label correlations across tasks, which improve performance for underrepresented classes. We also offer an extensive review of additional techniques proposed to address resource-poor problems in NLP, and show that none of these approaches can improve classification accuracy in our setting.

Low Frequency Names Exhibit Bias and Overfitting in Contextualizing Language Models
Robert Wolfe and Aylin Caliskan

We use a dataset of U.S. first names with labels based on predominant gender and racial group to examine the effect of training corpus frequency on tokenization, contextualization, similarity to initial representation, and bias in BERT, GPT-2, T5, and XLNet. We show that predominantly female and non-white names are less frequent in the training corpora of these four language models. We find that infrequent names are more self-similar across contexts, with Spearman's rho between frequency and self-similarity as low as -.763. Infrequent names are also less similar to initial representation, with Spearman's rho between frequency and linear centered kernel alignment (CKA) similarity to initial representation as high as .702. Moreover, we find Spearman's rho between racial bias and name frequency in BERT of .492, indicating that lower-frequency minority group names are more associated with unpleasantness. Representations of infrequent names undergo more processing, but are more self-similar, indicating that models rely on less context-informed representations of uncommon and minority names which are overfit to a lower number of observed contexts.

Mitigating Language-Dependent Ethnic Bias in BERT
Jaimeen Ahn and Alice Oh

In this paper, we study ethnic bias and how it varies across languages by analyzing and mitigating ethnic bias in monolingual BERT for English, German, Spanish, Korean, Turkish, and Chinese. To observe and quantify ethnic bias, we develop a novel metric called Categorical Bias score. Then we propose two methods for mitigation; first using a multilingual model, and second using contextual word alignment of two monolingual models. We compare our proposed methods with monolingual BERT and show that these methods effectively alleviate the ethnic bias. Which of the two methods works better depends on the amount of NLP resources available for that language. We additionally experiment with Arabic and Greek to verify that our proposed methods work for a wider variety of languages.

Adversarial Scrubbing of Demographic Information for Text Classification
Somnath Basu Roy Chowdhury et al.

Contextual representations learned by language models can often encode undesirable attributes, like demographic associations of the users, while being trained for an unrelated target task. We aim to scrub such undesirable attributes and learn fair representations while maintaining performance on the target task. In this paper, we present an adversarial learning framework "Adversarial Scrubber" (AdS), to debias contextual representations. We perform theoretical analysis to show that our framework converges without leaking demographic information under certain conditions. We extend previous evaluation techniques by evaluating debiasing performance using Minimum Description Length (MDL) probing. Experimental evaluations on 8 datasets show that AdS generates representations with minimal information about demographic attributes while being maximally informative about the target task.

Open-domain clarification question generation without question examples
Julia White et al.

An overarching goal of natural language processing is to enable machines to communicate seamlessly with humans. However, natural language can be ambiguous or unclear. In cases of uncertainty, humans engage in an interactive process known as repair: asking questions and seeking clarification until their uncertainty is resolved. We propose a framework for building a visually grounded question-asking model capable of producing polar (yes-no) clarification questions to resolve misunderstandings in dialogue. Our model uses an expected information gain objective to derive informative questions from an off-the-shelf image captioner without requiring any supervised question-answer data. We demonstrate our model's ability to pose questions that improve communicative success in a goal-oriented 20 questions game with synthetic and human answerers.

Improving Sequence-to-Sequence Pre-training via Sequence Span Rewriting
Wangchunshu Zhou et al.

In this paper, we propose **S**equence **S**pan **R**ewriting (SSR), a self-supervised task for sequence-to-sequence (Seq2Seq) pre-training. SSR learns to refine the machine-generated imperfect text spans into ground truth text. SSR provides more fine-grained and informative supervision in addition to the original text-infilling objective. Compared to the prevalent text infilling objectives for Seq2Seq pre-training, SSR is naturally more consistent with many downstream generation tasks that require sentence rewriting (e.g., text summarization, question generation, grammatical error correction, and paraphrase generation). We conduct extensive experiments by using SSR to improve the typical Seq2Seq pre-trained model T5 in a continual pre-training setting and show substantial improvements over T5 on various natural language generation tasks.

Coarse2Fine: Fine-grained Text Classification on Coarsely-grained Annotated Data
Dheeraj Mekala, Varun Gangal, and Jingbo Shang

Existing text classification methods mainly focus on a fixed label set, whereas many real-world applications require extending to new fine-grained classes as the number of samples per label increases. To accommodate such requirements, we introduce a new problem called coarse-to-fine grained classification, which aims to perform fine-grained classification on coarsely annotated data. Instead of asking for new fine-grained human annotations, we opt to leverage label surface names as the only human guidance and weave in rich pre-trained generative language models into the iterative weak supervision strategy. Specifically, we first propose a label-conditioned fine-tuning formulation to attune these generators for our task. Furthermore, we devise a regularization objective based on the coarse-fine label constraints derived from our problem setting, giving us even further improvements over the prior formulation. Our framework uses the fine-tuned generative models to sample pseudo-training data for training the classifier, and bootstraps on real unlabeled data for model refinement. Extensive experiments and case studies on two real-world datasets demonstrate superior performance over SOTA zero-shot classification baselines.

Text2Mol: Cross-Modal Molecule Retrieval with Natural Language Queries
Carl Edwards, ChengXiang Zhai, and Heng Ji

We propose a new task, Text2Mol, to retrieve molecules using natural language descriptions as queries. Natural language and molecules encode information in very different ways, which leads to the exciting but challenging problem of integrating these two very different modalities. Although some work has been done on text-based retrieval and structure-based retrieval, this new task requires integrating molecules and natural language more directly. Moreover, this can be viewed as an especially challenging cross-lingual retrieval problem by considering the molecules as a language with a very unique grammar. We construct a paired dataset of molecules and their corresponding text descriptions, which we use to learn an aligned common semantic embedding space for retrieval. We extend this to create a cross-modal attention-based model for explainability and reranking by interpreting the attentions as association rules. We also employ an ensemble approach to integrate our different architectures, which significantly improves results from 0.372 to 0.499 MRR. This new multimodal approach opens a new perspective on solving problems in chemistry literature understanding and molecular machine learning.

Neural Path Hunter: Reducing Hallucination in Dialogue Systems via Path Grounding
Nouha Dziri et al.

Dialogue systems powered by large pre-trained language models exhibit an innate ability to deliver fluent and natural-sounding responses. Despite their impressive performance, these models are fitful and can often generate factually incorrect statements impeding their widespread adoption. In this paper, we focus on the task of improving faithfulness and reducing hallucination of neural dialogue systems to known facts supplied by a Knowledge Graph (KG). We propose Neural Path Hunter which follows a generate-then-refine strategy whereby a generated response is amended using the KG. Neural Path Hunter leverages a separate token-level fact critic to identify plausible sources of hallucination followed by a refinement stage that retrieves correct entities by crafting a query signal that is propagated over a k-hop subgraph. We empirically validate our proposed approach on the OpenDialKG dataset (Moon et al., 2019) against a suite of metrics and report a relative improvement of faithfulness over dialogue responses by 20.35% based on FeQA (Durmus et al., 2020). The code is available at https://github.com/nouhadziri/Neural-Path-Hunter.

DuRecDial 2.0: A Bilingual Parallel Corpus for Conversational Recommendation
Zeming Liu et al.

In this paper, we provide a bilingual parallel human-to-human recommendation dialog dataset (DuRecDial 2.0) to enable researchers to explore a challenging task of multilingual and cross-lingual conversational recommendation. The difference between DuRecDial 2.0 and existing conversational recommendation datasets is that the data item (Profile, Goal, Knowledge, Context, Response) in DuRecDial 2.0 is annotated in two languages, both English and Chinese, while other datasets are built with the setting of a single language. We collect 8.2k dialogs aligned across English and Chinese languages (16.5k dialogs and 255k utterances in total) that are annotated by crowdsourced workers with strict quality control procedure. We then build monolingual, multilingual, and cross-lingual conversational recommendation baselines on DuRecDial 2.0. Experiment results show that the use of additional English data can bring performance improvement for Chinese conversational recommendation, indicating the benefits of DuRecDial 2.0. Finally, this dataset provides a challenging testbed for future studies of monolingual, multilingual, and cross-lingual conversational recommendation.

MindCraft: Theory of Mind Modeling for Situated Dialogue in Collaborative Tasks
Cristian-Paul Bara, Sky CH-Wang, and Joyce Chai

An ideal integration of autonomous agents in a human world implies that they are able to collaborate on human terms. In particular, theory of mind plays an important role in maintaining common ground during human collaboration and communication. To enable theory of mind modeling in situated interactions, we introduce a fine-grained dataset of collaborative tasks performed by pairs of human subjects in the 3D virtual blocks world of Minecraft. It provides information that captures partners' beliefs of the world and of each other as an interaction unfolds, bringing abundant opportunities to study human collaborative behaviors in situated language communication. As a first step towards our goal of developing embodied AI agents able to infer belief states of collaborative partners in situ, we build and present results on computational models for several theory of mind tasks.

Contextual Rephrase Detection for Reducing Friction in Dialogue Systems
Zhuoyi Wang et al.

For voice assistants like Alexa, Google Assistant, and Siri, correctly interpreting users' intentions is of utmost importance. However, users sometimes experience friction with these assistants, caused by errors from different system components or user errors such as slips of the tongue. Users tend to rephrase their queries until they get a satisfactory response. Rephrase detection is used to identify the rephrases and has long been treated as a task with pairwise input, which does not fully utilize the contextual information (e.g. users' implicit feedback). To this end, we propose a contextual rephrase detection model ContReph to automatically identify rephrases from multi-turn dialogues. We showcase how to leverage the dialogue context and user-agent interaction signals, including the user's implicit feedback and the time gap between different turns, which can help significantly outperform the pairwise rephrase detection models.

Understanding Politics via Contextualized Discourse Processing
Rajkumar Pujari and Dan Goldwasser

Politicians often have underlying agendas when reacting to events. Arguments in contexts of various events reflect a fairly consistent set of agendas for a given entity. In spite of recent advances in Pretrained Language Models, those text representations are not designed to capture such nuanced patterns. In this paper, we propose a Compositional Reader model consisting of encoder and composer modules, that captures and leverages such information to generate more effective representations for entities, issues, and events. These representations are contextualized by tweets, press releases, issues, news articles, and participating entities. Our model processes several documents at once and generates composed representations for multiple entities over several issues or events. Via qualitative and quantitative empirical analysis, we show that these representations are meaningful and effective.

Structural Adapters in Pretrained Language Models for AMR-to-Text Generation
Leonardo F. R. Ribeiro, Yue Zhang, and Iryna Gurevych

Pretrained language models (PLM) have recently advanced graph-to-text generation, where the input graph is linearized into a sequence and fed into the PLM to obtain its representation. However, efficiently encoding the graph structure in PLMs is challenging because such models were pretrained on natural language, and modeling structured data may lead to catastrophic forgetting of distributional knowledge. In this paper, we propose StructAdapt, an adapter method to encode graph structure into PLMs. Contrary to prior work, StructAdapt effectively models interactions among the nodes based on the graph connectivity, only training graph structure-aware adapter parameters. In this way, we incorporate task-specific knowledge while maintaining the topological structure of the graph. We empirically show the benefits of explicitly encoding graph structure into PLMs using StructAdapt, outperforming the state of the art on two AMR-to-text datasets, training only 5.1% of the PLM parameters.

Building Adaptive Acceptability Classifiers for Neural NLG
Soumya Batra et al.

We propose a novel framework to train models to classify acceptability of responses generated by natural language generation (NLG) models, improving upon existing sentence transformation and model-based approaches. An NLG response is considered acceptable if it is both semantically correct and grammatical. We don't make use of any human references making the classifiers suitable for runtime deployment. Training data for the classifiers is obtained using a 2-stage approach of first generating synthetic data using a combination of existing and new model-based approaches followed by a novel validation framework to filter and sort the synthetic data into acceptable and unacceptable classes. Our 2-stage approach adapts to a wide range of data representations and does not require additional data beyond what the NLG models are trained on. It is also independent of the underlying NLG model architecture, and is able to generate more realistic samples close to the distribution of the NLG model-generated responses. We present results on 5 datasets (WebNLG, Cleaned E2E, ViGGO, Alarm, and Weather) with varying data representations. We compare our framework with existing techniques that involve synthetic data generation using simple sentence transformations and/or model-based techniques, and show that building acceptability classifiers using data that resembles the generation model outputs followed by a validation framework outperforms the existing techniques, achieving state-of-the-art results. We also show that our techniques can be used in few-shot settings using self-training.

Continual Learning for Grounded Instruction Generation by Observing Human Following Behavior
Noriyuki Kojima, Alane Suhr, and Yoav Artzi

We study continual learning for natural language instruction generation, by observing human users' instruction execution. We focus on a collaborative scenario, where the system both acts and delegates tasks to human users using natural language. We compare user execution of generated instructions to the original system intent as an indication to the system's success communicating its intent. We show how to use this signal to improve the system's ability to generate instructions via contextual bandit learning. In interaction with real users, our system demonstrates dramatic improvements in its ability to generate language over time.

IR like a SIR: Sense-enhanced Information Retrieval for Multiple Languages
Rexhina Blloshmi et al.

With the advent of contextualized embeddings, attention towards neural ranking approaches for Information Retrieval increased considerably. However, two aspects have remained largely neglected: i) queries usually consist of few keywords only, which increases ambiguity and makes their contextualization harder, and ii) performing neural ranking on non-English documents is still cumbersome due to shortage of labeled datasets. In this paper we present SIR (Sense-enhanced Information Retrieval) to mitigate both problems by leveraging word sense information. At the core of our approach lies a novel multilingual query expansion mechanism based on Word Sense Disambiguation that provides sense definitions as additional semantic information for the query. Importantly, we use senses as a bridge across languages, thus allowing our model to perform considerably better than its supervised and unsupervised alternatives across French, German, Italian and Spanish languages on several CLEF benchmarks, while being trained on English Robust04 data only. We release SIR at https://github.com/SapienzaNLP/sir.

In-person Demo Session

Bavaro 4 *10:30–12:00*

COMBO: State-of-the-Art Morphosyntactic Analysis
Mateusz Klimaszewski and Alina Wróblewska

We introduce COMBO, a fully neural NLP system for accurate part-of-speech tagging, morphological analysis, lemmatisation, and (enhanced) dependency parsing. It predicts categorical morphosyntactic features whilst also exposes their vector representations, extracted from hidden layers. COMBO is an easy to install Python package with automatically downloadable pre-trained models for over 40 languages. It maintains a balance between efficiency and quality. As it is an end-to-end system and its modules are jointly trained, its training is competitively fast. As its models are optimised for accuracy, they achieve often better prediction quality than SOTA. The COMBO library is available at: https://gitlab.clarin-pl.eu/syntactic-tools/combo.

Summary Explorer: Visualizing the State of the Art in Text Summarization
Shahbaz Syed et al.

This paper introduces Summary Explorer, a new tool to support the manual inspection of text summarization systems by compiling the outputs of 55 state-of-the-art single document summarization approaches on three benchmark datasets, and visually exploring them during a qualitative assessment. The underlying design of the tool considers three well-known summary quality criteria (coverage, faithfulness, and position bias), encapsulated in a guided assessment based on tailored visualizations. The tool complements existing approaches for locally debugging summarization models and improves upon them. The tool is available at https://tldr.webis.de/

Session 2 Overview – Sunday, November 7, 2021

	Track A	Track B	Track C	Track D	Track E	Track F
	Machine Learning for NLP 1	*Generation 1*	*Interpretability and Analysis of Models for NLP 1*	*NLP Applications 1*	*Linguistic Theories, Cognitive Modeling and Psycholinguistics*	*Information Retrieval and Text Mining*
1:00	Classification of hierarchical text using geometric deep learning: the case of clinical trials corpus *Ferdowsi et al.*	Building Adaptive Acceptability Classifiers for Neural NLG *Batra et al.*	The Impact of Positional Encodings on Multilingual Compression *Ravishankar and Søgaard*	Memory and Knowledge Augmented Language Models for Inferring Salience in Long-Form Stories *Wilmot and Keller*	Predicting emergent linguistic compositions through time: Syntactic frame extension via multimodal chaining *Yu and Xu*	Condenser: a Pre-training Architecture for Dense Retrieval *Gao and Callan*
1:15	The Devil is in the Detail: Simple Tricks Improve Systematic Generalization of Transformers *Csordás, Irie, and Schmidhuber*	Moral Stories: Situated Reasoning about Norms, Intents, Actions, and their Consequences *Emelin et al.*	Disentangling Representations of Text by Masking Transformers *Zhang, Meent, and Wallace*	Semantic Novelty Detection in Natural Language Descriptions *Ma et al.*	Frequency Effects on Syntactic Rule Learning in Transformers *Wei et al.*	Monitoring geometrical properties of word embeddings for detecting the emergence of new topics. *Christophe et al.*
1:30	Artificial Text Detection via Examining the Topology of Attention Maps *Kushnareva et al.*	Continual Learning for Grounded Instruction Generation by Observing Human Following Behavior *Kojima, Suhr, and Artzi*	Exploring the Role of BERT Token Representations to Explain Sentence Probing Results *Mohebbi, Modarressi, and Pilehvar*	Jump-Starting Item Parameters for Adaptive Language Tests *McCarthy et al.*	A surprisal–duration trade-off across and within the world's languages *Pimentel et al.*	Contextualized Query Embeddings for Conversational Search *Lin, Yang, and Lin*

Track A	**Track B**	**Track C**	**Track D**	**Track E**	**Track F**	
Machine Learning for NLP 1	*Generation 1*	*Interpretability and Analysis of Models for NLP 1*	*NLP Applications 1*	*Linguistic Theories, Cognitive Modeling and Psycholinguistics*	*Information Retrieval and Text Mining*	
Active Learning by Acquiring Contrastive Examples *Margatina et al.*	Truth-Conditional Captions for Time Series Data *Jhamtani and Berg-Kirkpatrick*	Do Long-Range Language Models Actually Use Long-Range Context? *Sun et al.*	Lexically-Aware Semi-Supervised Learning for OCR Post-Correction *Rijhwani et al.*	Revisiting the Uniform Information Density Hypothesis *Meister et al.*	Ultra-High Dimensional Sparse Representations with Binarization for Efficient Text Retrieval *Jang et al.*	1:45
Conditional Poisson Stochastic Beams *Meister et al.*	Injecting Entity Types into Entity-Guided Text Generation *Dong et al.*	The World of an Octopus: How Reporting Bias Influences a Language Model's Perception of Color *Paik et al.*	Voice Query Auto Completion *Tang et al.*	The Taxonomy of Writing Systems: How to Measure how Logographic a System is *Sproat and Gutkin*	IR like a SIR: Sense-enhanced Information Retrieval for Multiple Languages *Blloshmi et al.*	2:00
Differentiable Subset Pruning of Transformer Heads *Li, Cotterell, and Sachan*	Smelting Gold and Silver for Improved Multilingual AMR-to-Text Generation *Ribeiro et al.*	SELFEXPLAIN: A Self-Explaining Architecture for Neural Text Classifiers *Rajagopal et al.*	CoPHE: A Count-Preserving Hierarchical Evaluation Metric in Large-Scale Multi-Label Text Classification *Falis et al.*	On the Relationships Between the Grammatical Genders of Inanimate Nouns and Their Co-Occurring Adjectives and Verbs *Williams et al.*	Neural Attention-Aware Hierarchical Topic Model *JIN et al.*	2:10
	Learning Compact Metrics for MT *Pu et al.*		Learning Universal Authorship Representations *Rivera-Soto et al.*			2:20

Parallel Session 2

Session 2A: Machine Learning for NLP 1

Bavaro 1 *Chair: Iz Beltagy*

Classification of hierarchical text using geometric deep learning: the case of clinical trials corpus

Sohrab Ferdowsi et al. 13:00–13:15

We consider the hierarchical representation of documents as graphs and use geometric deep learning to classify them into different categories. While graph neural networks can efficiently handle the variable structure of hierarchical documents using the permutation invariant message passing operations, we show that we can gain extra performance improvements using our proposed selective graph pooling operation that arises from the fact that some parts of the hierarchy are invariable across different documents. We applied our model to classify clinical trial (CT) protocols into completed and terminated categories. We use bag-of-words based, as well as pre-trained transformer-based embeddings to featurize the graph nodes, achieving f1-scoresaround 0.85 on a publicly available large scale CT registry of around 360K protocols. We further demonstrate how the selective pooling can add insights into the CT termination status prediction. We make the source code and dataset splits accessible.

The Devil is in the Detail: Simple Tricks Improve Systematic Generalization of Transformers

Róbert Csordás, Kazuki Irie, and Juergen Schmidhuber 13:15–13:30

Recently, many datasets have been proposed to test the systematic generalization ability of neural networks. The companion baseline Transformers, typically trained with default hyper-parameters from standard tasks, are shown to fail dramatically. Here we demonstrate that by revisiting model configurations as basic as scaling of embeddings, early stopping, relative positional embedding, and Universal Transformer variants, we can drastically improve the performance of Transformers on systematic generalization. We report improvements on five popular datasets: SCAN, CFQ, PCFG, COGS, and Mathematics dataset. Our models improve accuracy from 50% to 85% on the PCFG productivity split, and from 35% to 81% on COGS. On SCAN, relative positional embedding largely mitigates the EOS decision problem (Newman et al., 2020), yielding 100% accuracy on the length split with a cutoff at 26. Importantly, performance differences between these models are typically invisible on the IID data split. This calls for proper generalization validation sets for developing neural networks that generalize systematically. We publicly release the code to reproduce our results.

Artificial Text Detection via Examining the Topology of Attention Maps

Laida Kushnareva et al. 13:30–13:45

The impressive capabilities of recent generative models to create texts that are challenging to distinguish from the human-written ones can be misused for generating fake news, product reviews, and even abusive content. Despite the prominent performance of existing methods for artificial text detection, they still lack interpretability and robustness towards unseen models. To this end, we propose three novel types of interpretable topological features for this task based on Topological Data Analysis (TDA) which is currently understudied in the field of NLP. We empirically show that the features derived from the BERT model outperform count- and neural-based baselines up to 10% on three common datasets, and tend to be the most robust towards unseen GPT-style generation models as opposed to existing methods. The probing analysis of the features reveals their sensitivity to the surface and syntactic properties. The results demonstrate that TDA is a promising line with respect to NLP tasks, specifically the ones that incorporate surface and structural information.

Active Learning by Acquiring Contrastive Examples

Katerina Margatina et al. 13:45–14:00

Common acquisition functions for active learning use either uncertainty or diversity sampling, aiming to select difficult and diverse data points from the pool of unlabeled data, respectively. In this work, leveraging the best of both worlds, we propose an acquisition function that opts for selecting contrastive examples, i.e. data points that are similar in the model feature space and yet the model outputs maximally different predictive likelihoods. We compare our approach, CAL (Contrastive Active Learning), with a diverse set of acquisition functions in four natural language understanding tasks and seven datasets. Our experiments show that CAL performs consistently better or equal than the best performing baseline across all tasks, on both in-domain and out-of-domain data. We also conduct an extensive ablation study of our method and we further analyze all actively acquired datasets showing that CAL achieves a better trade-off between uncertainty and diversity compared to other strategies.

Conditional Poisson Stochastic Beams

Clara Meister et al. 14:00–14:15

Beam search is the default decoding strategy for many sequence generation tasks in NLP. The set of approximate K-best items returned by the algorithm is a useful summary of the distribution for many applications; however, the candidates typically exhibit high overlap and may give a highly biased estimate for expectations under our model. These problems can be addressed by instead using stochastic decoding strategies. In this work, we propose a new method for turning beam search into a stochastic process: Conditional Poisson

stochastic beam search. Rather than taking the maximizing set at each iteration, we sample K candidates without replacement according to the conditional Poisson sampling design. We view this as a more natural alternative to Kool et al. (2019)'s stochastic beam search (SBS). Furthermore, we show how samples generated under the CPSBS design can be used to build consistent estimators and sample diverse sets from sequence models. In our experiments, we observe CPSBS produces lower variance and more efficient estimators than SBS, even showing improvements in high entropy settings.

Differentiable Subset Pruning of Transformer Heads

Jiaoda Li, Ryan Cotterell, and Mrinmaya Sachan 14:15–14:30

Multi-head attention, a collection of several attention mechanisms that independently attend to different parts of the input, is the key ingredient in the Transformer (Vaswani et al., 2017). Recent work has shown, however, that a large proportion of the heads in a Transformer's multi-head attention mechanism can be safely pruned away without significantly harming the performance of the model; such pruning leads to models that are noticeably smaller and faster in practice. Our work introduces a new head pruning technique that we term differentiable subset pruning. Intuitively, our method learns per-head importance variables and then enforces a user-specified hard constraint on the number of unpruned heads. The importance variables are learned via stochastic gradient descent. We conduct experiments on natural language inference and machine translation; we show that differentiable subset pruning significantly improves over the previously proposed head pruning techniques of Voita et al. (2019) and Michel et al. (2019).

Session 2B: Generation 1

Bavaro 2 *Chair: Mohit Iyyer*

Building Adaptive Acceptability Classifiers for Neural NLG

Soumya Batra et al. 13:00–13:15

We propose a novel framework to train models to classify acceptability of responses generated by natural language generation (NLG) models, improving upon existing sentence transformation and model-based approaches. An NLG response is considered acceptable if it is both semantically correct and grammatical. We don't make use of any human references making the classifiers suitable for runtime deployment. Training data for the classifiers is obtained using a 2-stage approach of first generating synthetic data using a combination of existing and new model-based approaches followed by a novel validation framework to filter and sort the synthetic data into acceptable and unacceptable classes. Our 2-stage approach adapts to a wide range of data representations and does not require additional data beyond what the NLG models are trained on. It is also independent of the underlying NLG model architecture, and is able to generate more realistic samples close to the distribution of the NLG model-generated responses. We present results on 5 datasets (WebNLG, Cleaned E2E, ViGGO, Alarm, and Weather) with varying data representations. We compare our framework with existing techniques that involve synthetic data generation using simple sentence transformations and/or model-based techniques, and show that building acceptability classifiers using data that resembles the generation model outputs followed by a validation framework outperforms the existing techniques, achieving state-of-the-art results. We also show that our techniques can be used in few-shot settings using self-training.

Moral Stories: Situated Reasoning about Norms, Intents, Actions, and their Consequences

Denis Emelin et al. 13:15–13:30

In social settings, much of human behavior is governed by unspoken rules of conduct rooted in societal norms. For artificial systems to be fully integrated into social environments, adherence to such norms is a central prerequisite. To investigate whether language generation models can serve as behavioral priors for systems deployed in social settings, we evaluate their ability to generate action descriptions that achieve predefined goals under normative constraints. Moreover, we examine if models can anticipate likely consequences of actions that either observe or violate known norms, or explain why certain actions are preferable by generating relevant norm hypotheses. For this purpose, we introduce Moral Stories, a crowd-sourced dataset of structured, branching narratives for the study of grounded, goal-oriented social reasoning. Finally, we propose decoding strategies that combine multiple expert models to significantly improve the quality of generated actions, consequences, and norms compared to strong baselines.

Continual Learning for Grounded Instruction Generation by Observing Human Following Behavior

Noriyuki Kojima, Alane Suhr, and Yoav Artzi 13:30–13:45

We study continual learning for natural language instruction generation, by observing human users' instruction execution. We focus on a collaborative scenario, where the system both acts and delegates tasks to human users using natural language. We compare user execution of generated instructions to the original system intent as an indication to the system's success communicating its intent. We show how to use this signal to improve the system's ability to generate instructions via contextual bandit learning. In interaction with real users, our system demonstrates dramatic improvements in its ability to generate language over time.

Truth-Conditional Captions for Time Series Data

Harsh Jhamtani and Taylor Berg-Kirkpatrick 13:45–14:00

In this paper, we explore the task of automatically generating natural language descriptions of salient patterns in a time series, such as stock prices of a company over a week. A model for this task should be able to extract high-level patterns such as presence of a peak or a dip. While typical contemporary neural models with attention mechanisms can generate fluent output descriptions for this task, they often generate factually incorrect descriptions. We propose a computational model with a truth-conditional architecture which first runs small learned programs on the input time series, then identifies the programs/patterns which hold true for the given input, and finally conditions on *only* the chosen valid program (rather than the input time series) to generate the output text description. A program in our model is constructed from modules, which are small neural networks that are designed to capture numerical patterns and temporal information. The modules are shared across multiple programs, enabling compositionality as well as efficient learning of module parameters. The modules, as well as the composition of the modules, are unobserved in data, and we learn them in an end-to-end fashion with the only training signal coming from the accompanying natural language text descriptions. We find that the proposed model is able to generate high-precision captions even though we consider a small and simple space of module types.

Injecting Entity Types into Entity-Guided Text Generation

Xiangyu Dong et al. 14:00–14:10

Recent successes in deep generative modeling have led to significant advances in natural language generation (NLG). Incorporating entities into neural generation models has demonstrated great improvements by assisting to infer the summary topic and to generate coherent content. To enhance the role of entity in NLG, in this paper, we aim to model the entity type in the decoding phase to generate contextual words accurately. We develop a novel NLG model to produce a target sequence based on a given list of entities. Our model has a multi-

step decoder that injects the entity types into the process of entity mention generation. Experiments on two public news datasets demonstrate type injection performs better than existing type embedding concatenation baselines.

Smelting Gold and Silver for Improved Multilingual AMR-to-Text Generation
Leonardo F. R. Ribeiro et al. 14:10–14:20

Recent work on multilingual AMR-to-text generation has exclusively focused on data augmentation strategies that utilize silver AMR. However, this assumes a high quality of generated AMRs, potentially limiting the transferability to the target task. In this paper, we investigate different techniques for automatically generating AMR annotations, where we aim to study which source of information yields better multilingual results. Our models trained on gold AMR with silver (machine translated) sentences outperform approaches which leverage generated silver AMR. We find that combining both complementary sources of information further improves multilingual AMR-to-text generation. Our models surpass the previous state of the art for German, Italian, Spanish, and Chinese by a large margin.

Learning Compact Metrics for MT
Amy Pu et al. 14:20–14:30

Recent developments in machine translation and multilingual text generation have led researchers to adopt trained metrics such as COMET or BLEURT, which treat evaluation as a regression problem and use representations from multilingual pre-trained models such as XLM-RoBERTa or mBERT. Yet studies on related tasks suggest that these models are most efficient when they are large, which is costly and impractical for evaluation. We investigate the trade-off between multilinguality and model capacity with RemBERT, a state-of-the-art multilingual language model, using data from the WMT Metrics Shared Task. We present a series of experiments which show that model size is indeed a bottleneck for cross-lingual transfer, then demonstrate how distillation can help addressing this bottleneck, by leveraging synthetic data generation and transferring knowledge from one teacher to multiple students trained on related languages. Our method yields up to 10.5% improvement over vanilla fine-tuning and reaches 92.6% of RemBERT's performance using only a third of its parameters.

Session 2C: Interpretability and Analysis of Models for NLP 1

Bavaro 3 *Chair: Piyawat Lertvittayakumjorn*

The Impact of Positional Encodings on Multilingual Compression

Vinit Ravishankar and Anders Søgaard 13:00–13:15

In order to preserve word-order information in a non-autoregressive setting, transformer architectures tend to include positional knowledge, by (for instance) adding positional encodings to token embeddings. Several modifications have been proposed over the sinusoidal positional encodings used in the original transformer architecture; these include, for instance, separating position encodings and token embeddings, or directly modifying attention weights based on the distance between word pairs. We first show that surprisingly, while these modifications tend to improve monolingual language models, none of them result in better multilingual language models. We then answer why that is: sinusoidal encodings were explicitly designed to facilitate compositionality by allowing linear projections over arbitrary time steps. Higher variances in multilingual training distributions requires higher compression, in which case, compositionality becomes indispensable. Learned absolute positional encodings (e.g., in mBERT) tend to approximate sinusoidal embeddings in multilingual settings, but more complex positional encoding architectures lack the inductive bias to effectively learn cross-lingual alignment. In other words, while sinusoidal positional encodings were designed for monolingual applications, they are particularly useful in multilingual language models.

Disentangling Representations of Text by Masking Transformers

Xiongyi Zhang, Jan-Willem van de Meent, and Byron Wallace 13:15–13:30

Representations from large pretrained models such as BERT encode a range of features into monolithic vectors, affording strong predictive accuracy across a range of downstream tasks. In this paper we explore whether it is possible to learn disentangled representations by identifying existing subnetworks within pretrained models that encode distinct, complementary aspects. Concretely, we learn binary masks over transformer weights or hidden units to uncover subsets of features that correlate with a specific factor of variation; this eliminates the need to train a disentangled model from scratch for a particular task. We evaluate this method with respect to its ability to disentangle representations of sentiment from genre in movie reviews, toxicity from dialect in Tweets, and syntax from semantics. By combining masking with magnitude pruning we find that we can identify sparse subnetworks within BERT that strongly encode particular aspects (e.g., semantics) while only weakly encoding others (e.g., syntax). Moreover, despite only learning masks, disentanglement-via-masking performs as well as — and often better than —previously proposed methods based on variational autoencoders and adversarial training.

Exploring the Role of BERT Token Representations to Explain Sentence Probing Results

Hosein Mohebbi, Ali Modarressi, and Mohammad Taher Pilehvar 13:30–13:45

Several studies have been carried out on revealing linguistic features captured by BERT. This is usually achieved by training a diagnostic classifier on the representations obtained from different layers of BERT. The subsequent classification accuracy is then interpreted as the ability of the model in encoding the corresponding linguistic property. Despite providing insights, these studies have left out the potential role of token representations. In this paper, we provide a more in-depth analysis on the representation space of BERT in search for distinct and meaningful subspaces that can explain the reasons behind these probing results. Based on a set of probing tasks and with the help of attribution methods we show that BERT tends to encode meaningful knowledge in specific token representations (which are often ignored in standard classification setups), allowing the model to detect syntactic and semantic abnormalities, and to distinctively separate grammatical number and tense subspaces.

Do Long-Range Language Models Actually Use Long-Range Context?

Simeng Sun et al. 13:45–14:00

Language models are generally trained on short, truncated input sequences, which limits their ability to use discourse-level information present in long-range context to improve their predictions. Recent efforts to improve the efficiency of self-attention have led to a proliferation of long-range Transformer language models, which can process much longer sequences than models of the past. However, the ways in which such models take advantage of the long-range context remain unclear. In this paper, we perform a fine-grained analysis of two long-range Transformer language models (including the Routing Transformer, which achieves state-of-the-art perplexity on the PG-19 long-sequence LM benchmark dataset) that accept input sequences of up to 8K tokens. Our results reveal that providing long-range context (i.e., beyond the previous 2K tokens) to these models only improves their predictions on a small set of tokens (e.g., those that can be copied from the distant context) and does not help at all for sentence-level prediction tasks. Finally, we discover that PG-19 contains a variety of different document types and domains, and that long-range context helps most for literary novels (as opposed to textbooks or magazines).

The World of an Octopus: How Reporting Bias Influences a Language Model's Perception of Color

Cory Paik et al. 14:00–14:15

Recent work has raised concerns about the inherent limitations of text-only pretraining. In this paper, we first demonstrate that reporting bias, the tendency of people to not state the obvious, is one of the causes of this limitation, and then investigate to what extent multimodal training can mitigate this issue. To accomplish this,

we 1) generate the Color Dataset (CoDa), a dataset of human-perceived color distributions for 521 common objects; 2) use CoDa to analyze and compare the color distribution found in text, the distribution captured by language models, and a human's perception of color; and 3) investigate the performance differences between text-only and multimodal models on CoDa. Our results show that the distribution of colors that a language model recovers correlates more strongly with the inaccurate distribution found in text than with the ground-truth, supporting the claim that reporting bias negatively impacts and inherently limits text-only training. We then demonstrate that multimodal models can leverage their visual training to mitigate these effects, providing a promising avenue for future research.

SELFEXPLAIN: A Self-Explaining Architecture for Neural Text Classifiers
Dheeraj Rajagopal et al. 14:15–14:30

We introduce SelfExplain, a novel self-explaining model that explains a text classifier's predictions using phrase-based concepts. SelfExplain augments existing neural classifiers by adding (1) a globally interpretable layer that identifies the most influential concepts in the training set for a given sample and (2) a locally interpretable layer that quantifies the contribution of each local input concept by computing a relevance score relative to the predicted label. Experiments across five text-classification datasets show that SelfExplain facilitates interpretability without sacrificing performance. Most importantly, explanations from SelfExplain show sufficiency for model predictions and are perceived as adequate, trustworthy and understandable by human judges compared to existing widely-used baselines.

Session 2D: NLP Applications 1

Punta Cana 1,2 *Chair: Chrysoula Zerva*

Memory and Knowledge Augmented Language Models for Inferring Salience in Long-Form Stories

David Wilmot and Frank Keller 13:00–13:15

Measuring event salience is essential in the understanding of stories. This paper takes a recent unsupervised method for salience detection derived from Barthes Cardinal Functions and theories of surprise and applies it to longer narrative forms. We improve the standard transformer language model by incorporating an external knowledgebase (derived from Retrieval Augmented Generation) and adding a memory mechanism to enhance performance on longer works. We use a novel approach to derive salience annotation using chapter-aligned summaries from the Shmoop corpus for classic literary works. Our evaluation against this data demonstrates that our salience detection model improves performance over and above a non-knowledgebase and memory augmented language model, both of which are crucial to this improvement.

Semantic Novelty Detection in Natural Language Descriptions

Nianzu Ma et al. 13:15–13:30

This paper proposes to study a fine-grained semantic novelty detection task, which can be illustrated with the following example. It is normal that a person walks a dog in the park, but if someone says "A man is walking a chicken in the park", it is novel. Given a set of natural language descriptions of normal scenes, we want to identify descriptions of novel scenes. We are not aware of any existing work that solves the problem. Although existing novelty or anomaly detection algorithms are applicable, since they are usually topic-based, they perform poorly on our fine-grained semantic novelty detection task. This paper proposes an effective model (called GAT-MA) to solve the problem and also contributes a new dataset. Experimental evaluation shows that GAT-MA outperforms 11 baselines by large margins.

Jump-Starting Item Parameters for Adaptive Language Tests

Arya D. McCarthy et al. 13:30–13:45

A challenge in designing high-stakes language assessments is calibrating the test item difficulties, either a priori or from limited pilot test data. While prior work has addressed 'cold start' estimation of item difficulties without piloting, we devise a multi-task generalized linear model with BERT features to jump-start these estimates, rapidly improving their quality with as few as 500 test-takers and a small sample of item exposures (6 each) from a large item bank (4,000 items). Our joint model provides a principled way to compare test-taker proficiency, item difficulty, and language proficiency frameworks like the Common European Framework of Reference (CEFR). This also enables new item difficulty estimates without piloting them first, which in turn limits item exposure and thus enhances test item security. Finally, using operational data from the Duolingo English Test, a high-stakes English proficiency test, we find that the difficulty estimates derived using this method correlate strongly with lexico-grammatical features that correlate with reading complexity.

Lexically-Aware Semi-Supervised Learning for OCR Post-Correction

Shruti Rijhwani et al. 13:45–14:00

Much of the existing linguistic data in many languages of the world is locked away in non-digitized books and documents. Optical character recognition (OCR) can be used to produce digitized text, and previous work has demonstrated the utility of neural post-correction methods that improve the results of general-purpose OCR systems on recognition of less-well-resourced languages. However, these methods rely on manually curated post-correction data, which are relatively scarce compared to the non-annotated raw images that need to be digitized. In this paper, we present a semi-supervised learning method that makes it possible to utilize these raw images to improve performance, specifically through the use of self-training, a technique where a model is iteratively trained on its own outputs. In addition, to enforce consistency in the recognized vocabulary, we introduce a lexically-aware decoding method that augments the neural post-correction model with a count-based language model constructed from the recognized texts, implemented using weighted finite-state automata (WFSA) for efficient and effective decoding. Results on four endangered languages demonstrate the utility of the proposed method, with relative error reductions of 15-29%, where we find the combination of self-training and lexically-aware decoding essential for achieving consistent improvements.

Voice Query Auto Completion

Raphael Tang et al. 14:00–14:10

Query auto completion (QAC) is the task of predicting a search engine user's final query from their intermediate, incomplete query. In this paper, we extend QAC to the streaming voice search setting, where automatic speech recognition systems produce intermediate transcriptions as users speak. Naively applying existing methods fails because the intermediate transcriptions often don't form prefixes or even substrings of the final transcription. To address this issue, we propose to condition QAC approaches on intermediate transcriptions to complete voice queries. We evaluate our models on a speech-enabled smart television with real-life voice search traffic, finding that this ASR-aware conditioning improves the completion quality. Our best method obtains an 18% relative improvement in mean reciprocal rank over previous methods.

CoPHE: A Count-Preserving Hierarchical Evaluation Metric in Large-Scale Multi-Label Text Classification
Matúš Falis et al. 14:10–14:20

Large-Scale Multi-Label Text Classification (LMTC) includes tasks with hierarchical label spaces, such as automatic assignment of ICD-9 codes to discharge summaries. Performance of models in prior art is evaluated with standard precision, recall, and F1 measures without regard for the rich hierarchical structure. In this work we argue for hierarchical evaluation of the predictions of neural LMTC models. With the example of the ICD-9 ontology we describe a structural issue in the representation of the structured label space in prior art, and propose an alternative representation based on the depth of the ontology. We propose a set of metrics for hierarchical evaluation using the depth-based representation. We compare the evaluation scores from the proposed metrics with previously used metrics on prior art LMTC models for ICD-9 coding in MIMIC-III. We also propose further avenues of research involving the proposed ontological representation.

Learning Universal Authorship Representations
Rafael A. Rivera-Soto et al. 14:20–14:30

Determining whether two documents were composed by the same author, also known as authorship verification, has traditionally been tackled using statistical methods. Recently, authorship representations learned using neural networks have been found to outperform alternatives, particularly in large-scale settings involving hundreds of thousands of authors. But do such representations learned in a particular domain transfer to other domains? Or are these representations inherently entangled with domain-specific features? To study these questions, we conduct the first large-scale study of cross-domain transfer for authorship verification considering zero-shot transfers involving three disparate domains: Amazon reviews, fanfiction short stories, and Reddit comments. We find that although a surprising degree of transfer is possible between certain domains, it is not so successful between others. We examine properties of these domains that influence generalization and propose simple but effective methods to improve transfer.

Session 2E: Linguistic Theories, Cognitive Modeling and Psycholinguistics

Punta Cana 3,4 *Chair: Diana Inkpen*

Predicting emergent linguistic compositions through time: Syntactic frame extension via multimodal chaining

Lei Yu and Yang Xu 13:00–13:15

Natural language relies on a finite lexicon to express an unbounded set of emerging ideas. One result of this tension is the formation of new compositions, such that existing linguistic units can be combined with emerging items into novel expressions. We develop a framework that exploits the cognitive mechanisms of chaining and multimodal knowledge to predict emergent compositional expressions through time. We present the syntactic frame extension model (SFEM) that draws on the theory of chaining and knowledge from "percept", "concept", and "language" to infer how verbs extend their frames to form new compositions with existing and novel nouns. We evaluate SFEM rigorously on the 1) modalities of knowledge and 2) categorization models of chaining, in a syntactically parsed English corpus over the past 150 years. We show that multimodal SFEM predicts newly emerged verb syntax and arguments substantially better than competing models using purely linguistic or unimodal knowledge. We find support for an exemplar view of chaining as opposed to a prototype view and reveal how the joint approach of multimodal chaining may be fundamental to the creation of literal and figurative language uses including metaphor and metonymy.

Frequency Effects on Syntactic Rule Learning in Transformers

Jason Wei et al. 13:15–13:30

Pre-trained language models perform well on a variety of linguistic tasks that require symbolic reasoning, raising the question of whether such models implicitly represent abstract symbols and rules. We investigate this question using the case study of BERT's performance on English subject–verb agreement. Unlike prior work, we train multiple instances of BERT from scratch, allowing us to perform a series of controlled interventions at pre-training time. We show that BERT often generalizes well to subject–verb pairs that never occurred in training, suggesting a degree of rule-governed behavior. We also find, however, that performance is heavily influenced by word frequency, with experiments showing that both the absolute frequency of a verb form, as well as the frequency relative to the alternate inflection, are causally implicated in the predictions BERT makes at inference time. Closer analysis of these frequency effects reveals that BERT's behavior is consistent with a system that correctly applies the SVA rule in general but struggles to overcome strong training priors and to estimate agreement features (singular vs. plural) on infrequent lexical items.

A surprisal–duration trade-off across and within the world's languages

Tiago Pimentel et al. 13:30–13:45

While there exist scores of natural languages, each with its unique features and idiosyncrasies, they all share a unifying theme: enabling human communication. We may thus reasonably predict that human cognition shapes how these languages evolve and are used. Assuming that the capacity to process information is roughly constant across human populations, we expect a surprisal–duration trade-off to arise both across and within languages. We analyse this trade-off using a corpus of 600 languages and, after controlling for several potential confounds, we find strong supporting evidence in both settings. Specifically, we find that, on average, phones are produced faster in languages where they are less surprising, and vice versa. Further, we confirm that more surprising phones are longer, on average, in 319 languages out of the 600. We thus conclude that there is strong evidence of a surprisal–duration trade-off in operation, both across and within the world's languages.

Revisiting the Uniform Information Density Hypothesis

Clara Meister et al. 13:45–14:00

The uniform information density (UID) hypothesis posits a preference among language users for utterances structured such that information is distributed uniformly across a signal. While its implications on language production have been well explored, the hypothesis potentially makes predictions about language comprehension and linguistic acceptability as well. Further, it is unclear how uniformity in a linguistic signal—or lack thereof—should be measured, and over which linguistic unit, e.g., the sentence or language level, this uniformity should hold. Here we investigate these facets of the UID hypothesis using reading time and acceptability data. While our reading time results are generally consistent with previous work, they are also consistent with a weakly super-linear effect of surprisal, which would be compatible with UID's predictions. For acceptability judgments, we find clearer evidence that non-uniformity in information density is predictive of lower acceptability. We then explore multiple operationalizations of UID, motivated by different interpretations of the original hypothesis, and analyze the scope over which the pressure towards uniformity is exerted. The explanatory power of a subset of the proposed operationalizations suggests that the strongest trend may be a regression towards a mean surprisal across the language, rather than the phrase, sentence, or document—a finding that supports a typical interpretation of UID, namely that it is the byproduct of language users maximizing the use of a (hypothetical) communication channel.

The Taxonomy of Writing Systems: How to Measure how Logographic a System is

Richard Sproat and Alexander Gutkin 14:00–14:15

Taxonomies of writing systems since Gelb (1952) have classified systems based on what the written symbols represent: if they represent words or morphemes, they are logographic; if syllables, syllabic; if segments, al-

phabetic; and so forth. Sproat (2000) and Rogers (2005) broke with tradition by splitting the logographic and phonographic aspects into two dimensions, with logography being graded rather than a categorical distinction. A system could be syllabic, and highly logographic; or alphabetic, and mostly non-logographic. This accords better with how writing systems actually work, but neither author proposed a method for measuring logography. In this article we propose a novel measure of the degree of logography that uses an attention-based sequence-to-sequence model trained to predict the spelling of a token from its pronunciation in context. In an ideal phonographic system, the model should need to attend to only the current token in order to compute how to spell it, and this would show in the attention matrix activations. In contrast, with a logographic system, where a given pronunciation might correspond to several different spellings, the model would need to attend to a broader context. The ratio of the activation outside the token and the total activation forms the basis of our measure. We compare this with a simple lexical measure, and an entropic measure, as well as several other neural models, and argue that on balance our attention-based measure accords best with intuition about how logographic various systems are. Our work provides the first quantifiable measure of the notion of logography that accords with linguistic intuition and, we argue, provides better insight into what this notion means.

On the Relationships Between the Grammatical Genders of Inanimate Nouns and Their Co-Occurring Adjectives and Verbs

Adina Williams et al. 14:15–14:30

We use large-scale corpora in six different gendered languages, along with tools from NLP and information theory, to test whether there is a relationship between the grammatical genders of inanimate nouns and the adjectives used to describe those nouns. For all six languages, we find that there is a statistically significant relationship. We also find that there are statistically significant relationships between the grammatical genders of inanimate nouns and the verbs that take those nouns as direct objects, as indirect objects, and as subjects. We defer a deeper investigation of these relationships for future work.

Session 2F: Information Retrieval and Text Mining

Santo Domingo *Chair: Thuy Vu*

Condenser: a Pre-training Architecture for Dense Retrieval

Luyu Gao and Jamie Callan 13:00–13:15

Pre-trained Transformer language models (LM) have become go-to text representation encoders. Prior research fine-tunes deep LMs to encode text sequences such as sentences and passages into single dense vector representations for efficient text comparison and retrieval. However, dense encoders require a lot of data and sophisticated techniques to effectively train and suffer in low data situations. This paper finds a key reason is that standard LMs' internal attention structure is not ready-to-use for dense encoders, which needs to aggregate text information into the dense representation. We propose to pre-train towards dense encoder with a novel Transformer architecture, Condenser, where LM prediction CONditions on DENSE Representation. Our experiments show Condenser improves over standard LM by large margins on various text retrieval and similarity tasks.

Monitoring geometrical properties of word embeddings for detecting the emergence of new topics.

Clément Christophe et al. 13:15–13:30

Slow emerging topic detection is a task between event detection, where we aggregate behaviors of different words on short period of time, and language evolution, where we monitor their long term evolution. In this work, we tackle the problem of early detection of slowly emerging new topics. To this end, we gather evidence of weak signals at the word level. We propose to monitor the behavior of words representation in an embedding space and use one of its geometrical properties to characterize the emergence of topics. As evaluation is typically hard for this kind of task, we present a framework for quantitative evaluation and show positive results that outperform state-of-the-art methods. Our method is evaluated on two public datasets of press and scientific articles.

Contextualized Query Embeddings for Conversational Search

Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin 13:30–13:45

This paper describes a compact and effective model for low-latency passage retrieval in conversational search based on learned dense representations. Prior to our work, the state-of-the-art approach uses a multi-stage pipeline comprising conversational query reformulation and information retrieval modules. Despite its effectiveness, such a pipeline often includes multiple neural models that require long inference times. In addition, independently optimizing each module ignores dependencies among them. To address these shortcomings, we propose to integrate conversational query reformulation directly into a dense retrieval model. To aid in this goal, we create a dataset with pseudo-relevance labels for conversational search to overcome the lack of training data and to explore different training strategies. We demonstrate that our model effectively rewrites conversational queries as dense representations in conversational search and open-domain question answering datasets. Finally, after observing that our model learns to adjust the L2 norm of query token embeddings, we leverage this property for hybrid retrieval and to support error analysis.

Ultra-High Dimensional Sparse Representations with Binarization for Efficient Text Retrieval

Kyoung-Rok Jang et al. 13:45–14:00

The semantic matching capabilities of neural information retrieval can ameliorate synonymy and polysemy problems of symbolic approaches. However, neural models' dense representations are more suitable for re-ranking, due to their inefficiency. Sparse representations, either in symbolic or latent form, are more efficient with an inverted index. Taking the merits of the sparse and dense representations, we propose an ultra-high dimensional (UHD) representation scheme equipped with directly controllable sparsity. UHD's large capacity and minimal noise and interference among the dimensions allow for binarized representations, which are highly efficient for storage and search. Also proposed is a bucketing method, where the embeddings from multiple layers of BERT are selected/merged to represent diverse linguistic aspects. We test our models with MS MARCO and TREC CAR, showing that our models outperforms other sparse models.

IR like a SIR: Sense-enhanced Information Retrieval for Multiple Languages

Rexhina Blloshmi et al. 14:00–14:15

With the advent of contextualized embeddings, attention towards neural ranking approaches for Information Retrieval increased considerably. However, two aspects have remained largely neglected: i) queries usually consist of few keywords only, which increases ambiguity and makes their contextualization harder, and ii) performing neural ranking on non-English documents is still cumbersome due to shortage of labeled datasets. In this paper we present SIR (Sense-enhanced Information Retrieval) to mitigate both problems by leveraging word sense information. At the core of our approach lies a novel multilingual query expansion mechanism based on Word Sense Disambiguation that provides sense definitions as additional semantic information for the query. Importantly, we use senses as a bridge across languages, thus allowing our model to perform considerably better than its supervised and unsupervised alternatives across French, German, Italian and Spanish languages on several CLEF benchmarks, while being trained on English Robust04 data only. We release SIR at https://github.com/SapienzaNLP/sir.

Neural Attention-Aware Hierarchical Topic Model

YUAN JIN et al. 14:15–14:30

Neural topic models (NTMs) apply deep neural networks to topic modelling. Despite their success, NTMs generally ignore two important aspects: (1) only document-level word count information is utilized for the training, while more fine-grained sentence-level information is ignored, and (2) external semantic knowledge regarding documents, sentences and words are not exploited for the training. To address these issues, we propose a variational autoencoder (VAE) NTM model that jointly reconstructs the sentence and document word counts using combinations of bag-of-words (BoW) topical embeddings and pre-trained semantic embeddings. The pre-trained embeddings are first transformed into a common latent topical space to align their semantics with the BoW embeddings. Our model also features hierarchical KL divergence to leverage embeddings of each document to regularize those of their sentences, paying more attention to semantically relevant sentences. Both quantitative and qualitative experiments have shown the efficacy of our model in 1) lowering the reconstruction errors at both the sentence and document levels, and 2) discovering more coherent topics from real-world datasets.

Session 3 Overview – Sunday, November 7, 2021

	Track A	Track B	Track C	Track D	Track E	Track F
	Machine Learning for NLP 2	*Dialogue and Interactive Systems 1*	*Information Extraction 2*	*Resources and Evaluation 1*	*Discourse and Pragmatics*	*Semantics 1*
2:45	Relational World Knowledge Representation in Contextual Language Models: A Review *Safavi and Koutra*	MindCraft: Theory of Mind Modeling for Situated Dialogue in Collaborative Tasks *Bara, CH-Wang, and Chai*	Label Verbalization and Entailment for Effective Zero and Few-Shot Relation Extraction *Sainz et al.*	A Large-Scale Dataset for Empathetic Response Generation *Welivita, Xie, and Pu*	Understanding Politics via Contextualized Discourse Processing *Pujari and Goldwasser*	Asking It All: Generating Contextualized Questions for any Semantic Role *Pyatkin et al.*
3:00	Certified Robustness to Programmable Transformations in LSTMs *Zhang, Albarghouthi, and D'Antoni*	Detecting Speaker Personas from Conversational Texts *Gu et al.*	Extend, don't rebuild: Phrasing conditional graph modification as autoregressive sequence labelling *Weber et al.*	The Perils of Using Mechanical Turk to Evaluate Open-Ended Text Generation *Karpinska, Akoury, and Iyyer*	Conundrums in Event Coreference Resolution: Making Sense of the State of the Art *Lu and Ng*	Fast, Effective, and Self-Supervised: Transforming Masked Language Models into Universal Lexical and Sentence Encoders *Liu et al.*
3:15	ReGen: Reinforcement Learning for Text and Knowledge Base Generation using Pretrained Language Models *Dognin et al.*	Cross-lingual Intermediate Fine-tuning improves Dialogue State Tracking *Moghe, Steedman, and Birch*	Zero-Shot Information Extraction as a Unified Text-to-Triple Translation *Wang et al.*	Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus *Dodge et al.*	Weakly supervised discourse segmentation for multiparty oral conversations *Gravellier et al.*	RuleBERT: Teaching Soft Rules to Pre-Trained Language Models *Saeed et al.*
3:30	Contrastive Out-of-Distribution Detection for Pretrained Transformers *Zhou, Liu, and Chen*	ConvFiT: Conversational Fine-Tuning of Pretrained Language Models *Vulić et al.*	Learning Logic Rules for Document-Level Relation Extraction *Ru et al.*	AfroMT: Pretraining Strategies and Reproducible Benchmarks for Translation of 8 African Languages *Reid et al.*	Narrative Embedding: Re-Contextualization Through Attention *Wilner, Woolridge, and Glick*	Stepmothers are mean and academics are pretentious: What do pretrained language models learn about you? *Choenni, Shutova, and Rooij*

Track A	**Track B**	**Track C**	**Track D**	**Track E**	**Track F**	
Machine Learning for NLP 2	*Dialogue and Interactive Systems 1*	*Information Extraction 2*	*Resources and Evaluation 1*	*Discourse and Pragmatics*	*Semantics 1*	
Detecting Local Insights from Global Labels: Supervised & Zero-Shot Sequence Labeling via a Convolutional Decomposition *Schmaltz*	We've had this conversation before: A Novel Approach to Measuring Dialog Similarity *Lavi et al.*	Generalizing Cross-Document Event Coreference Resolution Across Multiple Corpora *Bugert, Reimers, and Gurevych*	Evaluating the Evaluation Metrics for Style Transfer: A Case Study in Multilingual Formality Transfer *Briakou et al.*	Focus on what matters: Applying Discourse Coherence Theory to Cross Document Coreference *Held, Iter, and Jurafsky*	ConSeC: Word Sense Disambiguation as Continuous Sense Comprehension *Barba, Procopio, and Navigli*	3:45
Measuring and Improving Consistency in Pretrained Language Models *Elazar et al.*	Towards Incremental Transformers: An Empirical Analysis of Transformer Models for Incremental NLU *Kahardipraja, Madureira, and Schlangen*	Revisiting Few-shot Relation Classification: Evaluation Data and Classification Schemes *Sabo et al.*	MS-Mentions: Consistently Annotating Entity Mentions in Materials Science Procedural Text *O'Gorman et al.*	Salience-Aware Event Chain Modeling for Narrative Understanding *Zhang, Chen, and May*	Shortcutted Commonsense: Data Spuriousness in Deep Learning of Commonsense Reasoning *Branco et al.*	3:55
	Feedback Attribution for Counterfactual Bandit Learning in Multi-Domain Spoken Language Understanding *Falke and Lehnen*					4:05

Parallel Session 3

Session 3A: Machine Learning for NLP 2

Bavaro 1 *Chair: Lei Li*

Relational World Knowledge Representation in Contextual Language Models: A Review

Tara Safavi and Danai Koutra 14:45–15:00

Relational knowledge bases (KBs) are commonly used to represent world knowledge in machines. However, while advantageous for their high degree of precision and interpretability, KBs are usually organized according to manually-defined schemas, which limit their expressiveness and require significant human efforts to engineer and maintain. In this review, we take a natural language processing perspective to these limitations, examining how they may be addressed in part by training deep contextual language models (LMs) to internalize and express relational knowledge in more flexible forms. We propose to organize knowledge representation strategies in LMs by the level of KB supervision provided, from no KB supervision at all to entity- and relation-level supervision. Our contributions are threefold: (1) We provide a high-level, extensible taxonomy for knowledge representation in LMs; (2) Within our taxonomy, we highlight notable models, evaluation tasks, and findings, in order to provide an up-to-date review of current knowledge representation capabilities in LMs; and (3) We suggest future research directions that build upon the complementary aspects of LMs and KBs as knowledge representations.

Certified Robustness to Programmable Transformations in LSTMs

Yuhao Zhang, Aws Albarghouthi, and Loris D'Antoni 15:00–15:15

Deep neural networks for natural language processing are fragile in the face of adversarial examples—small input perturbations, like synonym substitution or word duplication, which cause a neural network to change its prediction. We present an approach to certifying the robustness of LSTMs (and extensions of LSTMs) and training models that can be efficiently certified. Our approach can certify robustness to intractably large perturbation spaces defined programmatically in a language of string transformations. Our evaluation shows that (1) our approach can train models that are more robust to combinations of string transformations than those produced using existing techniques; (2) our approach can show high certification accuracy of the resulting models.

ReGen: Reinforcement Learning for Text and Knowledge Base Generation using Pretrained Language Models

Pierre Dognin et al. 15:15–15:30

Automatic construction of relevant Knowledge Bases (KBs) from text, and generation of semantically meaningful text from KBs are both long-standing goals in Machine Learning. In this paper, we present ReGen, a bidirectional generation of text and graph leveraging Reinforcement Learning to improve performance. Graph linearization enables us to re-frame both tasks as a sequence to sequence generation problem regardless of the generative direction, which in turn allows the use of Reinforcement Learning for sequence training where the model itself is employed as its own critic leading to Self-Critical Sequence Training (SCST). We present an extensive investigation demonstrating that the use of RL via SCST benefits graph and text generation on WebNLG+ 2020 and TekGen datasets. Our system provides state-of-the-art results on WebNLG+ 2020 by significantly improving upon published results from the WebNLG 2020+ Challenge for both text-to-graph and graph-to-text generation tasks. More details at https://github.com/IBM/regen.

Contrastive Out-of-Distribution Detection for Pretrained Transformers

Wenxuan Zhou, Fangyu Liu, and Muhao Chen 15:30–15:45

Pretrained Transformers achieve remarkable performance when training and test data are from the same distribution. However, in real-world scenarios, the model often faces out-of-distribution (OOD) instances that can cause severe semantic shift problems at inference time. Therefore, in practice, a reliable model should identify such instances, and then either reject them during inference or pass them over to models that handle another distribution. In this paper, we develop an unsupervised OOD detection method, in which only the in-distribution (ID) data are used in training. We propose to fine-tune the Transformers with a contrastive loss, which improves the compactness of representations, such that OOD instances can be better differentiated from ID ones. These OOD instances can then be accurately detected using the Mahalanobis distance in the model's penultimate layer. We experiment with comprehensive settings and achieve near-perfect OOD detection performance, outperforming baselines drastically. We further investigate the rationales behind the improvement, finding that more compact representations through margin-based contrastive learning bring the improvement. We release our code to the community for future research.

Detecting Local Insights from Global Labels: Supervised & Zero-Shot Sequence Labeling via a Convolutional Decomposition

Allen Schmaltz 15:45–16:00

We propose a new, more actionable view of neural network interpretability and data analysis by leveraging the remarkable matching effectiveness of representations derived from deep networks, guided by an approach for class-conditional feature detection. The decomposition of the filter-n-gram interactions of a convolutional neural network (CNN) and a linear layer over a pre-trained deep network yields a strong binary sequence la-

beler, with flexibility in producing predictions at—and defining loss functions for—varying label granularities, from the fully supervised sequence labeling setting to the challenging zero-shot sequence labeling setting, in which we seek token-level predictions but only have document-level labels for training. From this sequence-labeling layer we derive dense representations of the input that can then be matched to instances from training, or a support set with known labels. Such introspection with inference-time decision rules provides a means, in some settings, of making local updates to the model by altering the labels or instances in the support set without re-training the full model. Finally, we construct a particular K-nearest neighbors (K-NN) model from matched exemplar representations that approximates the original model's predictions and is at least as effective a predictor with respect to the ground-truth labels. This additionally yields interpretable heuristics at the token level for determining when predictions are less likely to be reliable, and for screening input dissimilar to the support set. In effect, we show that we can transform the deep network into a simple weighting over exemplars and associated labels, yielding an introspectable—and modestly updatable—version of the original model.

Measuring and Improving Consistency in Pretrained Language Models
Yanai Elazar et al. 16:00–16:15

Consistency of a model — that is, the invariance of its behavior under meaning-preserving alternations in its input — is a highly desirable property in natural language processing. In this paper we study the question: Are Pretrained Language Models (PLMs) consistent with respect to factual knowledge? To this end, we create ParaRel, a high-quality resource of cloze-style query English paraphrases. It contains a total of 328 paraphrases for 38 relations. Using ParaRel, we show that the consistency of all PLMs we experiment with is poor – though with high variance between relations. Our analysis of the representational spaces of PLMs suggests that they have a poor structure and are currently not suitable for representing knowledge robustly. Finally, we propose a method for improving model consistency and experimentally demonstrate its effectiveness.

Session 3B: Dialogue and Interactive Systems 1

Bavaro 2 *Chair: Kevin Small*

MindCraft: Theory of Mind Modeling for Situated Dialogue in Collaborative Tasks

Cristian-Paul Bara, Sky CH-Wang, and Joyce Chai 14:45–15:00

An ideal integration of autonomous agents in a human world implies that they are able to collaborate on human terms. In particular, theory of mind plays an important role in maintaining common ground during human collaboration and communication. To enable theory of mind modeling in situated interactions, we introduce a fine-grained dataset of collaborative tasks performed by pairs of human subjects in the 3D virtual blocks world of Minecraft. It provides information that captures partners' beliefs of the world and of each other as an interaction unfolds, bringing abundant opportunities to study human collaborative behaviors in situated language communication. As a first step towards our goal of developing embodied AI agents able to infer belief states of collaborative partners in situ, we build and present results on computational models for several theory of mind tasks.

Detecting Speaker Personas from Conversational Texts

Jia-Chen Gu et al. 15:00–15:15

Personas are useful for dialogue response prediction. However, the personas used in current studies are predefined and hard to obtain before a conversation. To tackle this issue, we study a new task, named Speaker Persona Detection (SPD), which aims to detect speaker personas based on the plain conversational text. In this task, a best-matched persona is searched out from candidates given the conversational text. This is a many-to-many semantic matching task because both contexts and personas in SPD are composed of multiple sentences. The long-term dependency and the dynamic redundancy among these sentences increase the difficulty of this task. We build a dataset for SPD, dubbed as Persona Match on Persona-Chat (PMPC). Furthermore, we evaluate several baseline models and propose utterance-to-profile (U2P) matching networks for this task. The U2P models operate at a fine granularity which treat both contexts and personas as sets of multiple sequences. Then, each sequence pair is scored and an interpretable overall score is obtained for a context-persona pair through aggregation. Evaluation results show that the U2P models outperform their baseline counterparts significantly.

Cross-lingual Intermediate Fine-tuning improves Dialogue State Tracking

Nikita Moghe, Mark Steedman, and Alexandra Birch 15:15–15:30

Recent progress in task-oriented neural dialogue systems is largely focused on a handful of languages, as annotation of training data is tedious and expensive. Machine translation has been used to make systems multilingual, but this can introduce a pipeline of errors. Another promising solution is using cross-lingual transfer learning through pretrained multilingual models. Existing methods train multilingual models with additional code-mixed task data or refine the cross-lingual representations through parallel ontologies. In this work, we enhance the transfer learning process by intermediate fine-tuning of pretrained multilingual models, where the multilingual models are fine-tuned with different but related data and/or tasks. Specifically, we use parallel and conversational movie subtitles datasets to design cross-lingual intermediate tasks suitable for downstream dialogue tasks. We use only 200K lines of parallel data for intermediate fine-tuning which is already available for 1782 language pairs. We test our approach on the cross-lingual dialogue state tracking task for the parallel MultiWoZ (English -> Chinese, Chinese -> English) and Multilingual WoZ (English -> German, English -> Italian) datasets. We achieve impressive improvements (> 20% on joint goal accuracy) on the parallel MultiWoZ dataset and the Multilingual WoZ dataset over the vanilla baseline with only 10% of the target language task data and zero-shot setup respectively.

ConvFiT: Conversational Fine-Tuning of Pretrained Language Models

Ivan Vulić et al. 15:30–15:45

Transformer-based language models (LMs) pretrained on large text collections are proven to store a wealth of semantic knowledge. However, 1) they are not effective as sentence encoders when used off-the-shelf, and 2) thus typically lag behind conversationally pretrained (e.g., via response selection) encoders on conversational tasks such as intent detection (ID). In this work, we propose ConvFiT, a simple and efficient two-stage procedure which turns any pretrained LM into a universal conversational encoder (after Stage 1 ConvFiT-ing) and task-specialised sentence encoder (after Stage 2). We demonstrate that 1) full-blown conversational pretraining is not required, and that LMs can be quickly transformed into effective conversational encoders with much smaller amounts of unannotated data; 2) pretrained LMs can be fine-tuned into task-specialised sentence encoders, optimised for the fine-grained semantics of a particular task. Consequently, such specialised sentence encoders allow for treating ID as a simple semantic similarity task based on interpretable nearest neighbours retrieval. We validate the robustness and versatility of the ConvFiT framework with such similarity-based inference on the standard ID evaluation sets: ConvFiT-ed LMs achieve state-of-the-art ID performance across the board, with particular gains in the most challenging, few-shot setups.

We've had this conversation before: A Novel Approach to Measuring Dialog Similarity

Ofer Lavi et al. 15:45–15:55

Dialog is a core building block of human natural language interactions. It contains multi-party utterances used to convey information from one party to another in a dynamic and evolving manner. The ability to compare dialogs is beneficial in many real world use cases, such as conversation analytics for contact center

calls and virtual agent design. We propose a novel adaptation of the edit distance metric to the scenario of dialog similarity. Our approach takes into account various conversation aspects such as utterance semantics, conversation flow, and the participants. We evaluate this new approach and compare it to existing document similarity measures on two publicly available datasets. The results demonstrate that our method outperforms the other approaches in capturing dialog flow, and is better aligned with the human perception of conversation similarity.

Towards Incremental Transformers: An Empirical Analysis of Transformer Models for Incremental NLU

Patrick Kahardipraja, Brielen Madureira, and David Schlangen 15:55–16:05

Incremental processing allows interactive systems to respond based on partial inputs, which is a desirable property e.g. in dialogue agents. The currently popular Transformer architecture inherently processes sequences as a whole, abstracting away the notion of time. Recent work attempts to apply Transformers incrementally via restart-incrementality by repeatedly feeding, to an unchanged model, increasingly longer input prefixes to produce partial outputs. However, this approach is computationally costly and does not scale efficiently for long sequences. In parallel, we witness efforts to make Transformers more efficient, e.g. the Linear Transformer (LT) with a recurrence mechanism. In this work, we examine the feasibility of LT for incremental NLU in English. Our results show that the recurrent LT model has better incremental performance and faster inference speed compared to the standard Transformer and LT with restart-incrementality, at the cost of part of the non-incremental (full sequence) quality. We show that the performance drop can be mitigated by training the model to wait for right context before committing to an output and that training with input prefixes is beneficial for delivering correct partial outputs.

Feedback Attribution for Counterfactual Bandit Learning in Multi-Domain Spoken Language Understanding

Tobias Falke and Patrick Lehnen 16:05–16:15

With counterfactual bandit learning, models can be trained based on positive and negative feedback received for historical predictions, with no labeled data needed. Such feedback is often available in real-world dialog systems, however, the modularized architecture commonly used in large-scale systems prevents the direct application of such algorithms. In this paper, we study the feedback attribution problem that arises when using counterfactual bandit learning for multi-domain spoken language understanding. We introduce an experimental setup to simulate the problem on small-scale public datasets, propose attribution methods inspired by multi-agent reinforcement learning and evaluate them against multiple baselines. We find that while directly using overall feedback leads to disastrous performance, our proposed attribution methods can allow training competitive models from user feedback.

Session 3C: Information Extraction 2

Bavaro 3 *Chair: Gerasimos Lampouras*

Label Verbalization and Entailment for Effective Zero and Few-Shot Relation Extraction

Oscar Sainz et al. 14:45–15:00

Relation extraction systems require large amounts of labeled examples which are costly to annotate. In this work we reformulate relation extraction as an entailment task, with simple, hand-made, verbalizations of relations produced in less than 15 min per relation. The system relies on a pretrained textual entailment engine which is run as-is (no training examples, zero-shot) or further fine-tuned on labeled examples (few-shot or fully trained). In our experiments on TACRED we attain 63% F1 zero-shot, 69% with 16 examples per relation (17% points better than the best supervised system on the same conditions), and only 4 points short to the state-of-the-art (which uses 20 times more training data). We also show that the performance can be improved significantly with larger entailment models, up to 12 points in zero-shot, allowing to report the best results to date on TACRED when fully trained. The analysis shows that our few-shot systems are specially effective when discriminating between relations, and that the performance difference in low data regimes comes mainly from identifying no-relation cases.

Extend, don't rebuild: Phrasing conditional graph modification as autoregressive sequence labelling

Leon Weber et al. 15:00–15:15

Deriving and modifying graphs from natural language text has become a versatile basis technology for information extraction with applications in many subfields, such as semantic parsing or knowledge graph construction. A recent work used this technique for modifying scene graphs (He et al. 2020), by first encoding the original graph and then generating the modified one based on this encoding. In this work, we show that we can considerably increase performance on this problem by phrasing it as graph extension instead of graph generation. We propose the first model for the resulting graph extension problem based on autoregressive sequence labelling. On three scene graph modification data sets, this formulation leads to improvements in accuracy over the state-of-the-art between 13 and 24 percentage points. Furthermore, we introduce a novel data set from the biomedical domain which has much larger linguistic variability and more complex graphs than the scene graph modification data sets. For this data set, the state-of-the art fails to generalize, while our model can produce meaningful predictions.

Zero-Shot Information Extraction as a Unified Text-to-Triple Translation

Chenguang Wang et al. 15:15–15:30

We cast a suite of information extraction tasks into a text-to-triple translation framework. Instead of solving each task relying on task-specific datasets and models, we formalize the task as a translation between task-specific input text and output triples. By taking the task-specific input, we enable a task-agnostic translation by leveraging the latent knowledge that a pre-trained language model has about the task. We further demonstrate that a simple pre-training task of predicting which relational information corresponds to which input text is an effective way to produce task-specific outputs. This enables the zero-shot transfer of our framework to downstream tasks. We study the zero-shot performance of this framework on open information extraction (OIE2016, NYT, WEB, PENN), relation classification (FewRel and TACRED), and factual probe (Google-RE and T-REx). The model transfers non-trivially to most tasks and is often competitive with a fully supervised method without the need for any task-specific training. For instance, we significantly outperform the F1 score of the supervised open information extraction without needing to use its training set.

Learning Logic Rules for Document-Level Relation Extraction

Dongyu Ru et al. 15:30–15:45

Document-level relation extraction aims to identify relations between entities in a whole document. Prior efforts to capture long-range dependencies have relied heavily on implicitly powerful representations learned through (graph) neural networks, which makes the model less transparent. To tackle this challenge, in this paper, we propose LogiRE, a novel probabilistic model for document-level relation extraction by learning logic rules. LogiRE treats logic rules as latent variables and consists of two modules: a rule generator and a relation extractor. The rule generator is to generate logic rules potentially contributing to final predictions, and the relation extractor outputs final predictions based on the generated logic rules. Those two modules can be efficiently optimized with the expectation-maximization (EM) algorithm. By introducing logic rules into neural networks, LogiRE can explicitly capture long-range dependencies as well as enjoy better interpretation. Empirical results show that significantly outperforms several strong baselines in terms of relation performance and logical consistency. Our code is available at https://github.com/rudongyu/LogiRE.

Generalizing Cross-Document Event Coreference Resolution Across Multiple Corpora

Michael Bugert, Nils Reimers, and Iryna Gurevych 15:45–16:00

Cross-document event coreference resolution (CDCR) is an NLP task in which mentions of events need to be identified and clustered throughout a collection of documents. CDCR aims to benefit downstream multidocument applications, but despite recent progress on corpora and system development, downstream improvements from applying CDCR have not been shown yet. We make the observation that every CDCR system to date was developed, trained, and tested only on a single respective corpus. This raises strong concerns on their generalizability—a must-have for downstream applications where the magnitude of domains or event mentions is

likely to exceed those found in a curated corpus. To investigate this assumption, we define a uniform evaluation setup involving three CDCR corpora: ECB+, the Gun Violence Corpus, and the Football Coreference Corpus (which we reannotate on token level to make our analysis possible). We compare a corpus-independent, feature-based system against a recent neural system developed for ECB+. Although being inferior in absolute numbers, the feature-based system shows more consistent performance across all corpora whereas the neural system is hit-or-miss. Via model introspection, we find that the importance of event actions, event time, and so forth, for resolving coreference in practice varies greatly between the corpora. Additional analysis shows that several systems overfit on the structure of the ECB+ corpus. We conclude with recommendations on how to achieve generally applicable CDCR systems in the future—the most important being that evaluation on multiple CDCR corpora is strongly necessary. To facilitate future research, we release our dataset, annotation guidelines, and system implementation to the public.

Revisiting Few-shot Relation Classification: Evaluation Data and Classification Schemes
Ofer Sabo et al. 16:00–16:15

We explore Few-Shot Learning (FSL) for Relation Classification (RC). Focusing on the realistic scenario of FSL, in which a test instance might not belong to any of the target categories (none-of-the-above, aka NOTA), we first revisit the recent popular dataset structure for FSL, pointing out its unrealistic data distribution. To remedy this, we propose a novel methodology for deriving more realistic few-shot test data from available datasets for supervised RC, and apply it to the TACRED dataset. This yields a new challenging benchmark for FSL RC, on which state of the art models show poor performance. Next, we analyze classification schemes within the popular embedding-based nearest-neighbor approach for FSL, with respect to constraints they impose on the embedding space. Triggered by this analysis we propose a novel classification scheme, in which the NOTA category is represented as learned vectors, shown empirically to be an appealing option for FSL.

Session 3D: Resources and Evaluation 1

Punta Cana 1,2 *Chair: Jacob Eisenstein*

A Large-Scale Dataset for Empathetic Response Generation

Anuradha Welivita, Yubo Xie, and Pearl Pu 14:45–15:00

Recent development in NLP shows a strong trend towards refining pre-trained models with a domain-specific dataset. This is especially the case for response generation where emotion plays an important role. However, existing empathetic datasets remain small, delaying research efforts in this area, for example, the development of emotion-aware chatbots. One main technical challenge has been the cost of manually annotating dialogues with the right emotion labels. In this paper, we describe a large-scale silver dataset consisting of 1M dialogues annotated with 32 fine-grained emotions, eight empathetic response intents, and the Neutral category. To achieve this goal, we have developed a novel data curation pipeline starting with a small seed of manually annotated data and eventually scaling it to a satisfactory size. We compare its quality against a state-of-the-art gold dataset using both offline experiments and visual validation methods. The resultant procedure can be used to create similar datasets in the same domain as well as in other domains.

The Perils of Using Mechanical Turk to Evaluate Open-Ended Text Generation

Marzena Karpinska, Nader Akoury, and Mohit Iyyer 15:00–15:15

Recent text generation research has increasingly focused on open-ended domains such as story and poetry generation. Because models built for such tasks are difficult to evaluate automatically, most researchers in the space justify their modeling choices by collecting crowdsourced human judgments of text quality (e.g., Likert scores of coherence or grammaticality) from Amazon Mechanical Turk (AMT). In this paper, we first conduct a survey of 45 open-ended text generation papers and find that the vast majority of them fail to report crucial details about their AMT tasks, hindering reproducibility. We then run a series of story evaluation experiments with both AMT workers and English teachers and discover that even with strict qualification filters, AMT workers (unlike teachers) fail to distinguish between model-generated text and human-generated references. We show that AMT worker judgments improve when they are shown model-generated output alongside human-generated references, which enables the workers to better calibrate their ratings. Finally, interviews with the English teachers provide deeper insights into the challenges of the evaluation process, particularly when rating model-generated text.

Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus

Jesse Dodge et al. 15:15–15:30

Large language models have led to remarkable progress on many NLP tasks, and researchers are turning to ever-larger text corpora to train them. Some of the largest corpora available are made by scraping significant portions of the internet, and are frequently introduced with only minimal documentation. In this work we provide some of the first documentation for the Colossal Clean Crawled Corpus (C4; Raffel et al., 2020), a dataset created by applying a set of filters to a single snapshot of Common Crawl. We begin by investigating where the data came from, and find a significant amount of text from unexpected sources like patents and US military websites. Then we explore the content of the text itself, and find machine-generated text (e.g., from machine translation systems) and evaluation examples from other benchmark NLP datasets. To understand the impact of the filters applied to create this dataset, we evaluate the text that was removed, and show that blocklist filtering disproportionately removes text from and about minority individuals. Finally, we conclude with some recommendations for how to created and document web-scale datasets from a scrape of the internet.

AfroMT: Pretraining Strategies and Reproducible Benchmarks for Translation of 8 African Languages

Machel Reid et al. 15:30–15:45

Reproducible benchmarks are crucial in driving progress of machine translation research. However, existing machine translation benchmarks have been mostly limited to high-resource or well-represented languages. Despite an increasing interest in low-resource machine translation, there are no standardized reproducible benchmarks for many African languages, many of which are used by millions of speakers but have less digitized textual data. To tackle these challenges, we propose AfroMT, a standardized, clean, and reproducible machine translation benchmark for eight widely spoken African languages. We also develop a suite of analysis tools for system diagnosis taking into account the unique properties of these languages. Furthermore, we explore the newly considered case of low-resource focused pretraining and develop two novel data augmentation-based strategies, leveraging word-level alignment information and pseudo-monolingual data for pretraining multilingual sequence-to-sequence models. We demonstrate significant improvements when pretraining on 11 languages, with gains of up to 2 BLEU points over strong baselines. We also show gains of up to 12 BLEU points over cross-lingual transfer baselines in data-constrained scenarios. All code and pretrained models will be released as further steps towards larger reproducible benchmarks for African languages.

Evaluating the Evaluation Metrics for Style Transfer: A Case Study in Multilingual Formality Transfer

Eleftheria Briakou et al. 15:45–16:00

While the field of style transfer (ST) has been growing rapidly, it has been hampered by a lack of standardized practices for automatic evaluation. In this paper, we evaluate leading automatic metrics on the oft-researched

task of formality style transfer. Unlike previous evaluations, which focus solely on English, we expand our focus to Brazilian-Portuguese, French, and Italian, making this work the first multilingual evaluation of metrics in ST. We outline best practices for automatic evaluation in (formality) style transfer and identify several models that correlate well with human judgments and are robust across languages. We hope that this work will help accelerate development in ST, where human evaluation is often challenging to collect.

MS-Mentions: Consistently Annotating Entity Mentions in Materials Science Procedural Text

Tim O'Gorman et al. 16:00–16:15

Material science synthesis procedures are a promising domain for scientific NLP, as proper modeling of these recipes could provide insight into new ways of creating materials. However, a fundamental challenge in building information extraction models for material science synthesis procedures is getting accurate labels for the materials, operations, and other entities of those procedures. We present a new corpus of entity mention annotations over 595 Material Science synthesis procedural texts (157,488 tokens), which greatly expands the training data available for the Named Entity Recognition task. We outline a new label inventory designed to provide consistent annotations and a new annotation approach intended to maximize the consistency and annotation speed of domain experts. Inter-annotator agreement studies and baseline models trained upon the data suggest that the corpus provides high-quality annotations of these mention types. This corpus helps lay a foundation for future high-quality modeling of synthesis procedures.

Session 3E: Discourse and Pragmatics

Punta Cana 3,4 *Chair: Kristen Johnson*

Understanding Politics via Contextualized Discourse Processing

Rajkumar Pujari and Dan Goldwasser 14:45–15:00

Politicians often have underlying agendas when reacting to events. Arguments in contexts of various events reflect a fairly consistent set of agendas for a given entity. In spite of recent advances in Pretrained Language Models, those text representations are not designed to capture such nuanced patterns. In this paper, we propose a Compositional Reader model consisting of encoder and composer modules, that captures and leverages such information to generate more effective representations for entities, issues, and events. These representations are contextualized by tweets, press releases, issues, news articles, and participating entities. Our model processes several documents at once and generates composed representations for multiple entities over several issues or events. Via qualitative and quantitative empirical analysis, we show that these representations are meaningful and effective.

Conundrums in Event Coreference Resolution: Making Sense of the State of the Art

Jing Lu and Vincent Ng 15:00–15:15

Despite recent promising results on the application of span-based models for event reference interpretation, there is a lack of understanding of what has been improved. We present an empirical analysis of a state-of-the-art span-based event reference systems with the goal of providing the general NLP audience with a better understanding of the state of the art and reference researchers with directions for future research.

Weakly supervised discourse segmentation for multiparty oral conversations

Lila Gravellier et al. 15:15–15:30

Discourse segmentation, the first step of discourse analysis, has been shown to improve results for text summarization, translation and other NLP tasks. While segmentation models for written text tend to perform well, they are not directly applicable to spontaneous, oral conversation, which has linguistic features foreign to written text. Segmentation is less studied for this type of language, where annotated data is scarce, and existing corpora more heterogeneous. We develop a weak supervision approach to adapt, using minimal annotation, a state of the art discourse segmenter trained on written text to French conversation transcripts. Supervision is given by a latent model bootstrapped by manually defined heuristic rules that use linguistic and acoustic information. The resulting model improves the original segmenter, especially in contexts where information on speaker turns is lacking or noisy, gaining up to 13% in F-score. Evaluation is performed on data like those used to define our heuristic rules, but also on transcripts from two other corpora.

Narrative Embedding: Re-Contextualization Through Attention

Sean Wilner, Daniel Woolridge, and Madeleine Glick 15:30–15:45

Narrative analysis is becoming increasingly important for a number of linguistic tasks including summarization, knowledge extraction, and question answering. We present a novel approach for narrative event representation using attention to re-contextualize events across the whole story. Comparing to previous analysis we find an unexpected attachment of event semantics to predicate tokens within a popular transformer model. We test the utility of our approach on narrative completion prediction, achieving state of the art performance on Multiple Choice Narrative Cloze and scoring competitively on the Story Cloze Task.

Focus on what matters: Applying Discourse Coherence Theory to Cross Document Coreference

William Held, Dan Iter, and Dan Jurafsky 15:45–16:00

Performing event and entity coreference resolution across documents vastly increases the number of candidate mentions, making it intractable to do the full n^2 pairwise comparisons. Existing approaches simplify by considering coreference only within document clusters, but this fails to handle inter-cluster coreference, common in many applications. As a result cross-document coreference algorithms are rarely applied to downstream tasks. We draw on an insight from discourse coherence theory: potential coreferences are constrained by the reader's discourse focus. We model the entities/events in a reader's focus as a neighborhood within a learned latent embedding space which minimizes the distance between mentions and the centroids of their gold coreference clusters. We then use these neighborhoods to sample only hard negatives to train a fine-grained classifier on mention pairs and their local discourse features. Our approach achieves state-of-the-art results for both events and entities on the ECB+, Gun Violence, Football Coreference, and Cross-Domain Cross-Document Coreference corpora. Furthermore, training on multiple corpora improves average performance across all datasets by 17.2 F1 points, leading to a robust coreference resolution model that is now feasible to apply to downstream tasks.

Salience-Aware Event Chain Modeling for Narrative Understanding

Xiyang Zhang, Muhao Chen, and Jonathan May 16:00–16:15

Storytelling, whether via fables, news reports, documentaries, or memoirs, can be thought of as the communication of interesting and related events that, taken together, form a concrete process. It is desirable to extract the event chains that represent such processes. However, this extraction remains a challenging problem. We posit that this is due to the nature of the texts from which chains are discovered. Natural language text in-

terleaves a narrative of concrete, salient events with background information, contextualization, opinion, and other elements that are important for a variety of necessary discourse and pragmatics acts but are not part of the principal chain of events being communicated. We introduce methods for extracting this principal chain from natural language text, by filtering away non-salient events and supportive sentences. We demonstrate the effectiveness of our methods at isolating critical event chains by comparing their effect on downstream tasks. We show that by pre-training large language models on our extracted chains, we obtain improvements in two tasks that benefit from a clear understanding of event chains: narrative prediction and event-based temporal question answering. The demonstrated improvements and ablative studies confirm that our extraction method isolates critical event chains.

Session 3F: Semantics 1

Santo Domingo — *Chair: Samuel Bowman*

Asking It All: Generating Contextualized Questions for any Semantic Role

Valentina Pyatkin et al. 14:45–15:00

Asking questions about a situation is an inherent step towards understanding it. To this end, we introduce the task of role question generation, which, given a predicate mention and a passage, requires producing a set of questions asking about all possible semantic roles of the predicate. We develop a two-stage model for this task, which first produces a context-independent question prototype for each role and then revises it to be contextually appropriate for the passage. Unlike most existing approaches to question generation, our approach does not require conditioning on existing answers in the text. Instead, we condition on the type of information to inquire about, regardless of whether the answer appears explicitly in the text, could be inferred from it, or should be sought elsewhere. Our evaluation demonstrates that we generate diverse and well-formed questions for a large, broad-coverage ontology of predicates and roles.

Fast, Effective, and Self-Supervised: Transforming Masked Language Models into Universal Lexical and Sentence Encoders

Fangyu Liu et al. 15:00–15:15

Previous work has indicated that pretrained Masked Language Models (MLMs) are not effective as universal lexical and sentence encoders off-the-shelf, i.e., without further task-specific fine-tuning on NLI, sentence similarity, or paraphrasing tasks using annotated task data. In this work, we demonstrate that it is possible to turn MLMs into effective lexical and sentence encoders even without any additional data, relying simply on self-supervision. We propose an extremely simple, fast, and effective contrastive learning technique, termed Mirror-BERT, which converts MLMs (e.g., BERT and RoBERTa) into such encoders in 20-30 seconds with no access to additional external knowledge. Mirror-BERT relies on identical and slightly modified string pairs as positive (i.e., synonymous) fine-tuning examples, and aims to maximise their similarity during "identity fine-tuning". We report huge gains over off-the-shelf MLMs with Mirror-BERT both in lexical-level and in sentence-level tasks, across different domains and different languages. Notably, in sentence similarity (STS) and question-answer entailment (QNLI) tasks, our self-supervised Mirror-BERT model even matches the performance of the Sentence-BERT models from prior work which rely on annotated task data. Finally, we delve deeper into the inner workings of MLMs, and suggest some evidence on why this simple Mirror-BERT fine-tuning approach can yield effective universal lexical and sentence encoders.

RuleBERT: Teaching Soft Rules to Pre-Trained Language Models

Mohammed Saeed et al. 15:15–15:30

While pre-trained language models (PLMs) are the go-to solution to tackle many natural language processing problems, they are still very limited in their ability to capture and to use common-sense knowledge. In fact, even if information is available in the form of approximate (soft) logical rules, it is not clear how to transfer it to a PLM in order to improve its performance for deductive reasoning tasks. Here, we aim to bridge this gap by teaching PLMs how to reason with soft Horn rules. We introduce a classification task where, given facts and soft rules, the PLM should return a prediction with a probability for a given hypothesis. We release the first dataset for this task, and we propose a revised loss function that enables the PLM to learn how to predict precise probabilities for the task. Our evaluation results show that the resulting fine-tuned models achieve very high performance, even on logical rules that were unseen at training. Moreover, we demonstrate that logical notions expressed by the rules are transferred to the fine-tuned model, yielding state-of-the-art results on external datasets.

Stepmothers are mean and academics are pretentious: What do pretrained language models learn about you?

Rochelle Choenni, Ekaterina Shutova, and Robert van Rooij 15:30–15:45

In this paper, we investigate what types of stereotypical information are captured by pretrained language models. We present the first dataset comprising stereotypical attributes of a range of social groups and propose a method to elicit stereotypes encoded by pretrained language models in an unsupervised fashion. Moreover, we link the emergent stereotypes to their manifestation as basic emotions as a means to study their emotional effects in a more generalized manner. To demonstrate how our methods can be used to analyze emotion and stereotype shifts due to linguistic experience, we use fine-tuning on news sources as a case study. Our experiments expose how attitudes towards different social groups vary across models and how quickly emotions and stereotypes can shift at the fine-tuning stage.

ConSeC: Word Sense Disambiguation as Continuous Sense Comprehension

Edoardo Barba, Luigi Procopio, and Roberto Navigli 15:45–16:00

Supervised systems have nowadays become the standard recipe for Word Sense Disambiguation (WSD), with Transformer-based language models as their primary ingredient. However, while these systems have certainly attained unprecedented performances, virtually all of them operate under the constraining assumption that, given a context, each word can be disambiguated individually with no account of the other sense choices. To address this limitation and drop this assumption, we propose CONtinuous SEnse Comprehension (ConSeC), a novel approach to WSD: leveraging a recent re-framing of this task as a text extraction problem, we adapt it to our formulation and introduce a feedback loop strategy that allows the disambiguation of a target word to

be conditioned not only on its context but also on the explicit senses assigned to nearby words. We evaluate ConSeC and examine how its components lead it to surpass all its competitors and set a new state of the art on English WSD. We also explore how ConSeC fares in the cross-lingual setting, focusing on 8 languages with various degrees of resource availability, and report significant improvements over prior systems. We release our code at https://github.com/SapienzaNLP/consec.

Shortcutted Commonsense: Data Spuriousness in Deep Learning of Commonsense Reasoning

Ruben Branco et al. 16:00–16:15

Commonsense is a quintessential human capacity that has been a core challenge to Artificial Intelligence since its inception. Impressive results in Natural Language Processing tasks, including in commonsense reasoning, have consistently been achieved with Transformer neural language models, even matching or surpassing human performance in some benchmarks. Recently, some of these advances have been called into question: so called data artifacts in the training data have been made evident as spurious correlations and shallow shortcuts that in some cases are leveraging these outstanding results. In this paper we seek to further pursue this analysis into the realm of commonsense related language processing tasks. We undertake a study on different prominent benchmarks that involve commonsense reasoning, along a number of key stress experiments, thus seeking to gain insight on whether the models are learning transferable generalizations intrinsic to the problem at stake or just taking advantage of incidental shortcuts in the data items. The results obtained indicate that most datasets experimented with are problematic, with models resorting to non-robust features and appearing not to be learning and generalizing towards the overall tasks intended to be conveyed or exemplified by the datasets.

Poster and Demo session 2

In-person Poster Session 2

Bavaro 4 *14:45–16:15*

When differential privacy meets NLP: The devil is in the detail

Ivan Habernal

Differential privacy provides a formal approach to privacy of individuals. Applications of differential privacy in various scenarios, such as protecting users' original utterances, must satisfy certain mathematical properties. Our contribution is a formal analysis of ADePT, a differentially private auto-encoder for text rewriting (Krishna et al, 2021). ADePT achieves promising results on downstream tasks while providing tight privacy guarantees. Our proof reveals that ADePT is not differentially private, thus rendering the experimental results unsubstantiated. We also quantify the impact of the error in its private mechanism, showing that the true sensitivity is higher by at least factor 6 in an optimistic case of a very small encoder's dimension and that the amount of utterances that are not privatized could easily reach 100% of the entire dataset. Our intention is neither to criticize the authors, nor the peer-reviewing process, but rather point out that if differential privacy applications in NLP rely on formal guarantees, these should be outlined in full and put under detailed scrutiny.

Achieving Model Robustness through Discrete Adversarial Training

Maor Ivgi and Jonathan Berant

Discrete adversarial attacks are symbolic perturbations to a language input that preserve the output label but lead to a prediction error. While such attacks have been extensively explored for the purpose of evaluating model robustness, their utility for improving robustness has been limited to offline augmentation only. Concretely, given a trained model, attacks are used to generate perturbed (adversarial) examples, and the model is re-trained exactly once. In this work, we address this gap and leverage discrete attacks for online augmentation, where adversarial examples are generated at every training step, adapting to the changing nature of the model. We propose (i) a new discrete attack, based on best-first search, and (ii) random sampling attacks that unlike prior work are not based on expensive search-based procedures. Surprisingly, we find that random sampling leads to impressive gains in robustness, outperforming the commonly-used offline augmentation, while leading to a speedup at training time of ~10x. Furthermore, online augmentation with search-based attacks justifies the higher training cost, significantly improving robustness on three datasets. Last, we show that our new attack substantially improves robustness compared to prior methods.

Debiasing Methods in Natural Language Understanding Make Bias More Accessible

Michael Mendelson and Yonatan Belinkov

Model robustness to bias is often determined by the generalization on carefully designed out-of-distribution datasets. Recent debiasing methods in natural language understanding (NLU) improve performance on such datasets by pressuring models into making unbiased predictions. An underlying assumption behind such methods is that this also leads to the discovery of more robust features in the model's inner representations. We propose a general probing-based framework that allows for post-hoc interpretation of biases in language models, and use an information-theoretic approach to measure the extractability of certain biases from the model's representations. We experiment with several NLU datasets and known biases, and show that, counter-intuitively, the more a language model is pushed towards a debiased regime, the more bias is actually encoded in its inner representations.

Evaluating the Robustness of Neural Language Models to Input Perturbations

Milad Moradi and Matthias Samwald

High-performance neural language models have obtained state-of-the-art results on a wide range of Natural Language Processing (NLP) tasks. However, results for common benchmark datasets often do not reflect model reliability and robustness when applied to noisy, real-world data. In this study, we design and implement various types of character-level and word-level perturbation methods to simulate realistic scenarios in which input texts may be slightly noisy or different from the data distribution on which NLP systems were trained. Conducting comprehensive experiments on different NLP tasks, we investigate the ability of high-performance language models such as BERT, XLNet, RoBERTa, and ELMo in handling different types of input perturbations. The results suggest that language models are sensitive to input perturbations and their performance can decrease even when small changes are introduced. We highlight that models need to be further improved and that current benchmarks are not reflecting model robustness well. We argue that evaluations on perturbed inputs should routinely complement widely-used benchmarks in order to yield a more realistic understanding of NLP systems' robustness.

How much pretraining data do language models need to learn syntax?

Laura Pérez-Mayos, Miguel Ballesteros, and Leo Wanner

Transformers-based pretrained language models achieve outstanding results in many well-known NLU benchmarks. However, while pretraining methods are very convenient, they are expensive in terms of time and resources. This calls for a study of the impact of pretraining data size on the knowledge of the models. We explore this impact on the syntactic capabilities of RoBERTa, using models trained on incremental sizes of raw text data. First, we use syntactic structural probes to determine whether models pretrained on more data

encode a higher amount of syntactic information. Second, we perform a targeted syntactic evaluation to analyze the impact of pretraining data size on the syntactic generalization performance of the models. Third, we compare the performance of the different models on three downstream applications: part-of-speech tagging, dependency parsing and paraphrase identification. We complement our study with an analysis of the cost-benefit trade-off of training such models. Our experiments show that while models pretrained on more data encode more syntactic knowledge and perform better on downstream applications, they do not always offer a better performance across the different syntactic phenomena and come at a higher financial and environmental cost.

Contrastive Explanations for Model Interpretability
Alon Jacovi et al.

Contrastive explanations clarify why an event occurred in contrast to another. They are inherently intuitive to humans to both produce and comprehend. We propose a method to produce contrastive explanations in the latent space, via a projection of the input representation, such that only the features that differentiate two potential decisions are captured. Our modification allows model behavior to consider only contrastive reasoning, and uncover which aspects of the input are useful for and against particular decisions. Our contrastive explanations can additionally answer for which label, and against which alternative label, is a given input feature useful. We produce contrastive explanations via both high-level abstract concept attribution and low-level input token/span attribution for two NLP classification benchmarks. Our findings demonstrate the ability of label-contrastive explanations to provide fine-grained interpretability of model decisions.

On the Transferability of Adversarial Attacks against Neural Text Classifier
Liping Yuan et al.

Deep neural networks are vulnerable to adversarial attacks, where a small perturbation to an input alters the model prediction. In many cases, malicious inputs intentionally crafted for one model can fool another model. In this paper, we present the first study to systematically investigate the transferability of adversarial examples for text classification models and explore how various factors, including network architecture, tokenization scheme, word embedding, and model capacity, affect the transferability of adversarial examples. Based on these studies, we propose a genetic algorithm to find an ensemble of models that can be used to induce adversarial examples to fool almost all existing models. Such adversarial examples reflect the defects of the learning process and the data bias in the training set. Finally, we derive word replacement rules that can be used for model diagnostics from these adversarial examples.

Conditional probing: measuring usable information beyond a baseline
John Hewitt et al.

Probing experiments investigate the extent to which neural representations make properties—like part-of-speech—predictable. One suggests that a representation encodes a property if probing that representation produces higher accuracy than probing a baseline representation like non-contextual word embeddings. Instead of using baselines as a point of comparison, we're interested in measuring information that is contained in the representation but not in the baseline. For example, current methods can detect when a representation is more useful than the word identity (a baseline) for predicting part-of-speech; however, they cannot detect when the representation is predictive of just the aspects of part-of-speech not explainable by the word identity. In this work, we extend a theory of usable information called V-information and propose conditional probing, which explicitly conditions on the information in the baseline. In a case study, we find that after conditioning on non-contextual word embeddings, properties like part-of-speech are accessible at deeper layers of a network than previously thought.

GFST: Gender-Filtered Self-Training for More Accurate Gender in Translation
Prafulla Kumar Choubey et al.

Targeted evaluations have found that machine translation systems often output incorrect gender in translations, even when the gender is clear from context. Furthermore, these incorrectly gendered translations have the potential to reflect or amplify social biases. We propose gender-filtered self-training (GFST) to improve gender translation accuracy on unambiguously gendered inputs. Our GFST approach uses a source monolingual corpus and an initial model to generate gender-specific pseudo-parallel corpora which are then filtered and added to the training data. We evaluate GFST on translation from English into five languages, finding that it improves gender accuracy without damaging generic quality. We also show the viability of GFST on several experimental settings, including re-training from scratch, fine-tuning, controlling the gender balance of the data, forward translation, and back-translation.

"Wikily" Supervised Neural Translation Tailored to Cross-Lingual Tasks
Mohammad Sadegh Rasooli, Chris Callison-Burch, and Derry Tanti Wijaya

We present a simple but effective approach for leveraging Wikipedia for neural machine translation as well as cross-lingual tasks of image captioning and dependency parsing without using any direct supervision from external parallel data or supervised models in the target language. We show that first sentences and titles of linked Wikipedia pages, as well as cross-lingual image captions, are strong signals for a seed parallel data to extract bilingual dictionaries and cross-lingual word embeddings for mining parallel text from Wikipedia. Our final model achieves high BLEU scores that are close to or sometimes higher than strong *supervised* baselines in low-resource languages; e.g. supervised BLEU of 4.0 versus 12.1 from our model in English-to-Kazakh. Moreover, we tailor our *wikily* translation models to unsupervised image captioning, and cross-

lingual dependency parser transfer. In image captioning, we train a multi-tasking machine translation and image captioning pipeline for Arabic and English from which the Arabic training data is a *wikily* translation of the English captioning data. Our captioning results on Arabic are slightly *better* than that of its supervised model. In dependency parsing, we translate a large amount of monolingual text, and use it as an artificial training data in an *annotation projection* framework. We show that our model outperforms recent work on cross-lingual transfer of dependency parsers.

mT6: Multilingual Pretrained Text-to-Text Transformer with Translation Pairs
Zewen Chi et al.

Multilingual T5 pretrains a sequence-to-sequence model on massive monolingual texts, which has shown promising results on many cross-lingual tasks. In this paper, we improve multilingual text-to-text transfer Transformer with translation pairs (mT6). Specifically, we explore three cross-lingual text-to-text pre-training tasks, namely, machine translation, translation pair span corruption, and translation span corruption. In addition, we propose a partially non-autoregressive objective for text-to-text pre-training. We evaluate the methods on seven multilingual benchmark datasets, including sentence classification, named entity recognition, question answering, and abstractive summarization. Experimental results show that the proposed mT6 improves cross-lingual transferability over mT5.

Improving Zero-Shot Cross-Lingual Transfer Learning via Robust Training
Kuan-Hao Huang et al.

Pre-trained multilingual language encoders, such as multilingual BERT and XLM-R, show great potential for zero-shot cross-lingual transfer. However, these multilingual encoders do not precisely align words and phrases across languages. Especially, learning alignments in the multilingual embedding space usually requires sentence-level or word-level parallel corpora, which are expensive to be obtained for low-resource languages. An alternative is to make the multilingual encoders more robust; when fine-tuning the encoder using downstream task, we train the encoder to tolerate noise in the contextual embedding spaces such that even if the representations of different languages are not aligned well, the model can still achieve good performance on zero-shot cross-lingual transfer. In this work, we propose a learning strategy for training robust models by drawing connections between adversarial examples and the failure cases of zero-shot cross-lingual transfer. We adopt two widely used robust training methods, adversarial training and randomized smoothing, to train the desired robust model. The experimental results demonstrate that robust training improves zero-shot cross-lingual transfer on text classification tasks. The improvement is more significant in the generalized cross-lingual transfer setting, where the pair of input sentences belong to two different languages.

Speechformer: Reducing Information Loss in Direct Speech Translation
Sara Papi et al.

Transformer-based models have gained increasing popularity achieving state-of-the-art performance in many research fields including speech translation. However, Transformer's quadratic complexity with respect to the input sequence length prevents its adoption as is with audio signals, which are typically represented by long sequences. Current solutions resort to an initial sub-optimal compression based on a fixed sampling of raw audio features. Therefore, potentially useful linguistic information is not accessible to higher-level layers in the architecture. To solve this issue, we propose Speechformer, an architecture that, thanks to reduced memory usage in the attention layers, avoids the initial lossy compression and aggregates information only at a higher level according to more informed linguistic criteria. Experiments on three language pairs (en→de/es/nl) show the efficacy of our solution, with gains of up to 0.8 BLEU on the standard MuST-C corpus and of up to 4.0 BLEU in a low resource scenario.

Is "moby dick" a Whale or a Bird? Named Entities and Terminology in Speech Translation
Marco Gaido et al.

Automatic translation systems are known to struggle with rare words. Among these, named entities (NEs) and domain-specific terms are crucial, since errors in their translation can lead to severe meaning distortions. Despite their importance, previous speech translation (ST) studies have neglected them, also due to the dearth of publicly available resources tailored to their specific evaluation. To fill this gap, we i) present the first systematic analysis of the behavior of state-of-the-art ST systems in translating NEs and terminology, and ii) release NEuRoparl-ST, a novel benchmark built from European Parliament speeches annotated with NEs and terminology. Our experiments on the three language directions covered by our benchmark (en→es/fr/it) show that ST systems correctly translate 75—80% of terms and 65—70% of NEs, with very low performance (37—40%) on person names.

HintedBT: Augmenting Back-Translation with Quality and Transliteration Hints
Sahana Ramnath et al.

Back-translation (BT) of target monolingual corpora is a widely used data augmentation strategy for neural machine translation (NMT), especially for low-resource language pairs. To improve effectiveness of the available BT data, we introduce HintedBT—a family of techniques which provides hints (through tags) to the encoder and decoder. First, we propose a novel method of using both high and low quality BT data by providing hints (as source tags on the encoder) to the model about the quality of each source-target pair. We don't filter out low quality data but instead show that these hints enable the model to learn effectively from noisy data. Second, we address the problem of predicting whether a source token needs to be translated or transliterated to the target language, which is common in cross-script translation tasks (i.e., where source and target do not

share the written script). For such cases, we propose training the model with additional hints (as target tags on the decoder) that provide information about the operation required on the source (translation or both translation and transliteration). We conduct experiments and detailed analyses on standard WMT benchmarks for three cross-script low/medium-resource language pairs: Hindi,Gujarati,Tamil-to-English. Our methods compare favorably with five strong and well established baselines. We show that using these hints, both separately and together, significantly improves translation quality and leads to state-of-the-art performance in all three language pairs in corresponding bilingual settings.

Translation-based Supervision for Policy Generation in Simultaneous Neural Machine Translation
Ashkan Alinejad, Hassan S. Shavarani, and Anoop Sarkar

In simultaneous machine translation, finding an agent with the optimal action sequence of reads and writes that maintain a high level of translation quality while minimizing the average lag in producing target tokens remains an extremely challenging problem. We propose a novel supervised learning approach for training an agent that can detect the minimum number of reads required for generating each target token by comparing simultaneous translations against full-sentence translations during training to generate oracle action sequences. These oracle sequences can then be used to train a supervised model for action generation at inference time. Our approach provides an alternative to current heuristic methods in simultaneous translation by introducing a new training objective, which is easier to train than previous attempts at training the agent using reinforcement learning techniques for this task. Our experimental results show that our novel training method for action generation produces much higher quality translations while minimizing the average lag in simultaneous translation.

Nearest Neighbour Few-Shot Learning for Cross-lingual Classification
M Saiful Bari, Batool Haider, and Saab Mansour

Even though large pre-trained multilingual models (e.g. mBERT, XLM-R) have led to significant performance gains on a wide range of cross-lingual NLP tasks, success on many downstream tasks still relies on the availability of sufficient annotated data. Traditional fine-tuning of pre-trained models using only a few target samples can cause over-fitting. This can be quite limiting as most languages in the world are under-resourced. In this work, we investigate cross-lingual adaptation using a simple nearest-neighbor few-shot (<15 samples) inference technique for classification tasks. We experiment using a total of 16 distinct languages across two NLP tasks- XNLI and PAWS-X. Our approach consistently improves traditional fine-tuning using only a handful of labeled samples in target locales. We also demonstrate its generalization capability across tasks.

Cross-Attention is All You Need: Adapting Pretrained Transformers for Machine Translation
Mozhdeh Gheini, Xiang Ren, and Jonathan May

We study the power of cross-attention in the Transformer architecture within the context of transfer learning for machine translation, and extend the findings of studies into cross-attention when training from scratch. We conduct a series of experiments through fine-tuning a translation model on data where either the source or target language has changed. These experiments reveal that fine-tuning only the cross-attention parameters is nearly as effective as fine-tuning all parameters (i.e., the entire translation model). We provide insights into why this is the case and observe that limiting fine-tuning in this manner yields cross-lingually aligned embeddings. The implications of this finding for researchers and practitioners include a mitigation of catastrophic forgetting, the potential for zero-shot translation, and the ability to extend machine translation models to several new language pairs with reduced parameter storage overhead.

A Bayesian Framework for Information-Theoretic Probing
Tiago Pimentel and Ryan Cotterell

Pimentel et al. (2020) recently analysed probing from an information-theoretic perspective. They argue that probing should be seen as approximating a mutual information. This led to the rather unintuitive conclusion that representations encode exactly the same information about a target task as the original sentences. The mutual information, however, assumes the true probability distribution of a pair of random variables is known, leading to unintuitive results in settings where it is not. This paper proposes a new framework to measure what we term Bayesian mutual information, which analyses information from the perspective of Bayesian agents—allowing for more intuitive findings in scenarios with finite data. For instance, under Bayesian MI we have that data can add information, processing can help, and information can hurt, which makes it more intuitive for machine learning applications. Finally, we apply our framework to probing where we believe Bayesian mutual information naturally operationalises ease of extraction by explicitly limiting the available background knowledge to solve a task.

Measuring Association Between Labels and Free-Text Rationales
Sarah Wiegreffe, Ana Marasović, and Noah A. Smith

In interpretable NLP, we require faithful rationales that reflect the model's decision-making process for an explained instance. While prior work focuses on extractive rationales (a subset of the input words), we investigate their less-studied counterpart: free-text natural language rationales. We demonstrate that *pipelines*, models for faithful rationalization on information-extraction style tasks, do not work as well on "reasoning" tasks requiring free-text rationales. We turn to models that *jointly* predict and rationalize, a class of widely used high-performance models for free-text rationalization. We investigate the extent to which the labels and

rationales predicted by these models are associated, a necessary property of faithful explanation. Via two tests, *robustness equivalence* and *feature importance agreement*, we find that state-of-the-art T5-based joint models exhibit desirable properties for explaining commonsense question-answering and natural language inference, indicating their potential for producing faithful free-text rationales.

Are Transformers a Modern Version of ELIZA? Observations on French Object Verb Agreement
Bingzhi Li, Guillaume Wisniewski, and Benoit Crabbé

Many recent works have demonstrated that unsupervised sentence representations of neural networks encode syntactic information by observing that neural language models are able to predict the agreement between a verb and its subject. We take a critical look at this line of research by showing that it is possible to achieve high accuracy on this agreement task with simple surface heuristics, indicating a possible flaw in our assessment of neural networks' syntactic ability. Our fine-grained analyses of results on the long-range French object-verb agreement show that contrary to LSTMs, Transformers are able to capture a non-trivial amount of grammatical structure.

Do Long-Range Language Models Actually Use Long-Range Context?
Simeng Sun et al.

Language models are generally trained on short, truncated input sequences, which limits their ability to use discourse-level information present in long-range context to improve their predictions. Recent efforts to improve the efficiency of self-attention have led to a proliferation of long-range Transformer language models, which can process much longer sequences than models of the past. However, the ways in which such models take advantage of the long-range context remain unclear. In this paper, we perform a fine-grained analysis of two long-range Transformer language models (including the Routing Transformer, which achieves state-of-the-art perplexity on the PG-19 long-sequence LM benchmark dataset) that accept input sequences of up to 8K tokens. Our results reveal that providing long-range context (i.e., beyond the previous 2K tokens) to these models only improves their predictions on a small set of tokens (e.g., those that can be copied from the distant context) and does not help at all for sentence-level prediction tasks. Finally, we discover that PG-19 contains a variety of different document types and domains, and that long-range context helps most for literary novels (as opposed to textbooks or magazines).

The World of an Octopus: How Reporting Bias Influences a Language Model's Perception of Color
Cory Paik et al.

Recent work has raised concerns about the inherent limitations of text-only pretraining. In this paper, we first demonstrate that reporting bias, the tendency of people to not state the obvious, is one of the causes of this limitation, and then investigate to what extent multimodal training can mitigate this issue. To accomplish this, we 1) generate the Color Dataset (CoDa), a dataset of human-perceived color distributions for 521 common objects; 2) use CoDa to analyze and compare the color distribution found in text, the distribution captured by language models, and a human's perception of color; and 3) investigate the performance differences between text-only and multimodal models on CoDa. Our results show that the distribution of colors that a language model recovers correlates more strongly with the inaccurate distribution found in text than with the ground-truth, supporting the claim that reporting bias negatively impacts and inherently limits text-only training. We then demonstrate that multimodal models can leverage their visual training to mitigate these effects, providing a promising avenue for future research.

Studying word order through iterative shuffling
Nikolay Malkin et al.

As neural language models approach human performance on NLP benchmark tasks, their advances are widely seen as evidence of an increasingly complex understanding of syntax. This view rests upon a hypothesis that has not yet been empirically tested: that word order encodes meaning essential to performing these tasks. We refute this hypothesis in many cases: in the GLUE suite and in various genres of English text, the words in a sentence or phrase can rarely be permuted to form a phrase carrying substantially different information. Our surprising result relies on inference by iterative shuffling (IBIS), a novel, efficient procedure that finds the ordering of a bag of words having the highest likelihood under a fixed language model. IBIS can use any black-box model without additional training and is superior to existing word ordering algorithms. Coalescing our findings, we discuss how shuffling inference procedures such as IBIS can benefit language modeling and constrained generation.

SELFEXPLAIN: A Self-Explaining Architecture for Neural Text Classifiers
Dheeraj Rajagopal et al.

We introduce SelfExplain, a novel self-explaining model that explains a text classifier's predictions using phrase-based concepts. SelfExplain augments existing neural classifiers by adding (1) a globally interpretable layer that identifies the most influential concepts in the training set for a given sample and (2) a locally interpretable layer that quantifies the contribution of each local input concept by computing a relevance score relative to the predicted label. Experiments across five text-classification datasets show that SelfExplain facilitates interpretability without sacrificing performance. Most importantly, explanations from SelfExplain show sufficiency for model predictions and are perceived as adequate, trustworthy and understandable by human

judges compared to existing widely-used baselines.

Data and Parameter Scaling Laws for Neural Machine Translation
Mitchell A Gordon, Kevin Duh, and Jared Kaplan

We observe that the development cross-entropy loss of supervised neural machine translation models scales like a power law with the amount of training data and the number of non-embedding parameters in the model. We discuss some practical implications of these results, such as predicting BLEU achieved by large scale models and predicting the ROI of labeling data in low-resource language pairs.

On the Difficulty of Translating Free-Order Case-Marking Languages
Arianna Bisazza, Ahmet Üstün, and Stephan Sportel

Identifying factors that make certain languages harder to model than others is essential to reach language equality in future Natural Language Processing technologies. Free-order case-marking languages, such as Russian, Latin or Tamil, have proved more challenging than fixed-order languages for the tasks of syntactic parsing and subject-verb agreement prediction. In this work, we investigate whether this class of languages is also more difficult to translate by state-of-the-art Neural Machine Translation models (NMT). Using a variety of synthetic languages and a newly introduced translation challenge set, we find that word order flexibility in the source language only leads to a very small loss of NMT quality, even though the core verb arguments become impossible to disambiguate in sentences without semantic cues. The latter issue is indeed solved by the addition of case marking. However, in medium- and low-resource settings, the overall NMT quality of fixed-order languages remains unmatched.

Toward Deconfounding the Influence of Subject's Demographic Characteristics in Question Answering
Maharshi Gor, Kellie Webster, and Jordan Boyd-Graber

The goal of question answering (QA) is to answer _any_ question. However, major QA datasets have skewed distributions over gender, profession, and nationality. Despite that skew, an analysis of model accuracy reveals little evidence that accuracy is lower for people based on gender or nationality; instead, there is more variation on professions (question topic) and question ambiguity. But QA's lack of representation could itself hide evidence of bias, necessitating QA datasets that better represent global diversity.

In-person Demo Session

Bavaro 4 *14:45–16:15*

Thermostat: A Large Collection of NLP Model Explanations and Analysis Tools
Nils Feldhus, Robert Schwarzenberg, and Sebastian Möller

In the language domain, as in other domains, neural explainability takes an ever more important role, with feature attribution methods on the forefront. Many such methods require considerable computational resources and expert knowledge about implementation details and parameter choices. To facilitate research, we present Thermostat which consists of a large collection of model explanations and accompanying analysis tools. Thermostat allows easy access to over 200k explanations for the decisions of prominent state-of-the-art models spanning across different NLP tasks, generated with multiple explainers. The dataset took over 10k GPU hours (> one year) to compile; compute time that the community now saves. The accompanying software tools allow to analyse explanations instance-wise but also accumulatively on corpus level. Users can investigate and compare models, datasets and explainers without the need to orchestrate implementation details. Thermostat is fully open source, democratizes explainability research in the language domain, circumvents redundant computations and increases comparability and replicability.

Box Embeddings: An open-source library for representation learning using geometric structures
Tejas Chheda et al.

A fundamental component to the success of modern representation learning is the ease of performing various vector operations. Recently, objects with more geometric structure (eg. distributions, complex or hyperbolic vectors, or regions such as cones, disks, or boxes) have been explored for their alternative inductive biases and additional representational capacity. In this work, we introduce Box Embeddings, a Python library that enables researchers to easily apply and extend probabilistic box embeddings. Fundamental geometric operations on boxes are implemented in a numerically stable way, as are modern approaches to training boxes which mitigate gradient sparsity. The library is fully open source, and compatible with both PyTorch and TensorFlow, which allows existing neural network layers to be replaced with or transformed into boxes easily. In this work, we present the implementation details of the fundamental components of the library, and the concepts required to use box representations alongside existing neural network architectures.

Session 4 Overview – Sunday, November 7, 2021

	Track A	Track B	Track C	Track D	Track E	Track F
	Machine Learning for NLP 3	*Dialogue and Interactive Systems 2*	*Information Extraction 3*	*Ethics and NLP*	*Phonology, Morphology and Word Segmentation*	*Speech, Vision, Robotics, Multimodal Grounding 1*
4:45	Effects of Parameter Norm Growth During Transformer Training: Inductive Bias from Gradient Descent *Merrill et al.*	CR-Walker: Tree-Structured Graph Reasoning and Dialog Acts for Conversational Recommendation *Ma, Takanobu, and Huang*	AttentionRank: Unsupervised Keyphrase Extraction using Self and Cross Attentions *Ding and Luo*	Harms of Gender Exclusivity and Challenges in Non-Binary Representation in Language Technologies *Dev et al.*	Local Word Discovery for Interactive Transcription *Lane and Bird*	Broaden the Vision: Geo-Diverse Visual Commonsense Reasoning *Yin et al.*
5:00	Foreseeing the Benefits of Incidental Supervision *He et al.*	DIALKI: Knowledge Identification in Conversational Systems through Dialogue-Document Contextualization *Wu et al.*	Unsupervised Relation Extraction: A Variational Autoencoder Approach *Yuan and Eldardiry*	Are Gender-Neutral Queries Really Gender-Neutral? Mitigating Gender Bias in Image Search *Wang, Liu, and Wang*	Segment, Mask, and Predict: Augmenting Chinese Word Segmentation with Self-Supervision *Maimaiti et al.*	Reference-Centric Models for Grounded Collaborative Dialogue *Fried, Chiu, and Klein*
5:15	Competency Problems: On Finding and Removing Artifacts in Language Data *Gardner et al.*	Iconary: A Pictionary-Based Game for Testing Multimodal Communication with Drawings and Text *Clark et al.*	Robust Retrieval Augmented Generation for Zero-shot Slot Filling *Glass et al.*	Style Pooling: Automatic Text Style Obfuscation for Improved Classification Fairness *Mireshghallah and Berg-Kirkpatrick*	Minimal Supervision for Morphological Inflection *Goldman and Tsarfaty*	CrossVQA: Scalably Generating Benchmarks for Systematically Testing VQA Generalization *Akula et al.*
5:30	Sensitivity as a Complexity Measure for Sequence Classification Tasks *Hahn, Jurafsky, and Futrell*	Self-training Improves Pre-training for Few-shot Learning in Task-oriented Dialog Systems *Mi et al.*	Everything Is All It Takes: A Multipronged Strategy for Zero-Shot Cross-Lingual Information Extraction *Yarmohammadi et al.*	Modeling Disclosive Transparency in NLP Application Descriptions *Saxon et al.*	Fast WordPiece Tokenization *Song et al.*	Generative Spoken Language Modeling from Raw Audio *Lakhotia et al.*

Track A	**Track B**	**Track C**	**Track D**	**Track E**	**Track F**	
Machine Learning for NLP 3	*Dialogue and Interactive Systems 2*	*Information Extraction 3*	*Ethics and NLP*	*Phonology, Morphology and Word Segmentation*	*Speech, Vision, Robotics, Multimodal Grounding 1*	
Knowledge-Aware Meta-learning for Low-Resource Text Classification *Yao et al.*	Contextual Rephrase Detection for Reducing Friction in Dialogue Systems *Wang et al.*	Partially Supervised Named Entity Recognition via the Expected Entity Ratio Loss *Effland and Collins*	Reconstruction Attack on Instance Encoding for Language Understanding *Xie and Hong*	You should evaluate your language model on marginal likelihood over tokenisations *Cao and Rimell*	Visual Goal-Step Inference using wikiHow *Yang et al.*	5:45
Sentence Bottleneck Autoencoders from Transformer Language Models *Montero, Pappas, and Smith*	Few-Shot Intent Detection via Contrastive Pre-Training and Fine-Tuning *Zhang et al.*	MasakhaNER: Named Entity Recognition for African Languages *Abbott et al.*	Fairness-aware Class Imbalanced Learning *Subramanian et al.*	Identity-Based Patterns in Deep Convolutional Networks: Generative Adversarial Phonology and Reduplication *Beguš*	Systematic Generalization on gSCAN: What is Nearly Solved and What is Next? *Qiu et al.*	5:55
Efficient Contrastive Learning via Novel Data Augmentation and Curriculum Learning *Ye, Kim, and Oh*	"It doesn't look good for a date": Transforming Critiques into Preferences for Conversational Recommendation Systems *Bursztyn et al.*		CRYPTOGRU: Low Latency Privacy-Preserving Text Analysis With GRU *Feng et al.*		Effect of Visual Extensions on Natural Language Understanding in Vision-and-Language Models *Iki and Aizawa*	6:05

Parallel Session 4

Session 4A: Machine Learning for NLP 3

Bavaro 1 *Chair: Sachin Kumar*

Effects of Parameter Norm Growth During Transformer Training: Inductive Bias from Gradient Descent

William Merrill et al. 16:45–17:00

The capacity of neural networks like the widely adopted transformer is known to be very high. Evidence is emerging that they learn successfully due to inductive bias in the training routine, typically a variant of gradient descent (GD). To better understand this bias, we study the tendency for transformer parameters to grow in magnitude (ℓ_2 norm) during training, and its implications for the emergent representations within self attention layers. Empirically, we document norm growth in the training of transformer language models, including T5 during its pretraining. As the parameters grow in magnitude, we prove that the network approximates a discretized network with saturated activation functions. Such "saturated" networks are known to have a reduced capacity compared to the full network family that can be described in terms of formal languages and automata. Our results suggest saturation is a new characterization of an inductive bias implicit in GD of particular interest for NLP. We leverage the emergent discrete structure in a saturated transformer to analyze the role of different attention heads, finding that some focus locally on a small number of positions, while other heads compute global averages, allowing counting. We believe understanding the interplay between these two capabilities may shed further light on the structure of computation within large transformers.

Foreseeing the Benefits of Incidental Supervision

Hangfeng He et al. 17:00–17:15

Real-world applications often require improved models by leveraging *a range of cheap incidental supervision signals*. These could include partial labels, noisy labels, knowledge-based constraints, and cross-domain or cross-task annotations – all having statistical associations with gold annotations but not exactly the same. However, we currently lack a principled way to measure the benefits of these signals to a given target task, and the common practice of evaluating these benefits is through exhaustive experiments with various models and hyperparameters. This paper studies whether we can, *in a single framework, quantify the benefits of various types of incidental signals for a given target task without going through combinatorial experiments*. We propose a unified PAC-Bayesian motivated informativeness measure, PABI, that characterizes the uncertainty reduction provided by incidental supervision signals. We demonstrate PABI's effectiveness by quantifying the value added by various types of incidental signals to sequence tagging tasks. Experiments on named entity recognition (NER) and question answering (QA) show that PABI's predictions correlate well with learning performance, providing a promising way to determine, ahead of learning, which supervision signals would be beneficial.

Competency Problems: On Finding and Removing Artifacts in Language Data

Matt Gardner et al. 17:15–17:30

Much recent work in NLP has documented dataset artifacts, bias, and spurious correlations between input features and output labels. However, how to tell which features have "spurious" instead of legitimate correlations is typically left unspecified. In this work we argue that for complex language understanding tasks, all simple feature correlations are spurious, and we formalize this notion into a class of problems which we call competency problems. For example, the word "amazing" on its own should not give information about a sentiment label independent of the context in which it appears, which could include negation, metaphor, sarcasm, etc. We theoretically analyze the difficulty of creating data for competency problems when human bias is taken into account, showing that realistic datasets will increasingly deviate from competency problems as dataset size increases. This analysis gives us a simple statistical test for dataset artifacts, which we use to show more subtle biases than were described in prior work, including demonstrating that models are inappropriately affected by these less extreme biases. Our theoretical treatment of this problem also allows us to analyze proposed solutions, such as making local edits to dataset instances, and to give recommendations for future data collection and model design efforts that target competency problems.

Sensitivity as a Complexity Measure for Sequence Classification Tasks

Michael Hahn, Dan Jurafsky, and Richard Futrell 17:30–17:45

We introduce a theoretical framework for understanding and predicting the complexity of sequence classification tasks, using a novel extension of the theory of Boolean function sensitivity. The sensitivity of a function, given a distribution over input sequences, quantifies the number of disjoint subsets of the input sequence that can each be individually changed to change the output. We argue that standard sequence classification methods are biased towards learning low-sensitivity functions, so that tasks requiring high sensitivity are more difficult. To that end, we show analytically that simple lexical classifiers can only express functions of bounded sensitivity, and we show empirically that low-sensitivity functions are easier to learn for LSTMs. We then estimate sensitivity on 15 NLP tasks, finding that sensitivity is higher on challenging tasks collected in GLUE than on simple text classification tasks, and that sensitivity predicts the performance both of simple lexical classifiers and of vanilla BiLSTMs without pretrained contextualized embeddings. Within a task, sensitivity predicts which inputs are hard for such simple models. Our results suggest that the success of massively pretrained contextual representations stems in part because they provide representations from which information can be

extracted by low-sensitivity decoders.

Knowledge-Aware Meta-learning for Low-Resource Text Classification
Huaxiu Yao et al. 17:45–17:55

Meta-learning has achieved great success in leveraging the historical learned knowledge to facilitate the learning process of the new task. However, merely learning the knowledge from the historical tasks, adopted by current meta-learning algorithms, may not generalize well to testing tasks when they are not well-supported by training tasks. This paper studies a low-resource text classification problem and bridges the gap between meta-training and meta-testing tasks by leveraging the external knowledge bases. Specifically, we propose KGML to introduce additional representation for each sentence learned from the extracted sentence-specific knowledge graph. The extensive experiments on three datasets demonstrate the effectiveness of KGML under both supervised adaptation and unsupervised adaptation settings.

Sentence Bottleneck Autoencoders from Transformer Language Models
Ivan Montero, Nikolaos Pappas, and Noah A. Smith 17:55–18:05

Representation learning for text via pretraining a language model on a large corpus has become a standard starting point for building NLP systems. This approach stands in contrast to autoencoders, also trained on raw text, but with the objective of learning to encode each input as a vector that allows full reconstruction. Autoencoders are attractive because of their latent space structure and generative properties. We therefore explore the construction of a sentence-level autoencoder from a pretrained, frozen transformer language model. We adapt the masked language modeling objective as a generative, denoising one, while only training a sentence bottleneck and a single-layer modified transformer decoder. We demonstrate that the sentence representations discovered by our model achieve better quality than previous methods that extract representations from pretrained transformers on text similarity tasks, style transfer (an example of controlled generation), and single-sentence classification tasks in the GLUE benchmark, while using fewer parameters than large pretrained models.

Efficient Contrastive Learning via Novel Data Augmentation and Curriculum Learning
Seonghyeon Ye, Jiseon Kim, and Alice Oh 18:05–18:15

We introduce EfficientCL, a memory-efficient continual pretraining method that applies contrastive learning with novel data augmentation and curriculum learning. For data augmentation, we stack two types of operation sequentially: cutoff and PCA jittering. While pretraining steps proceed, we apply curriculum learning by incrementing the augmentation degree for each difficulty step. After data augmentation is finished, contrastive learning is applied on projected embeddings of original and augmented examples. When finetuned on GLUE benchmark, our model outperforms baseline models, especially for sentence-level tasks. Additionally, this improvement is capable with only 70% of computational memory compared to the baseline model.

Session 4B: Dialogue and Interactive Systems 2

Bavaro 2 *Chair: Alborz Geramifard*

CR-Walker: Tree-Structured Graph Reasoning and Dialog Acts for Conversational Recommendation

Wenchang Ma, Ryuichi Takanobu, and Minlie Huang 16:45–17:00

Growing interests have been attracted in Conversational Recommender Systems (CRS), which explore user preference through conversational interactions in order to make appropriate recommendation. However, there is still a lack of ability in existing CRS to (1) traverse multiple reasoning paths over background knowledge to introduce relevant items and attributes, and (2) arrange selected entities appropriately under current system intents to control response generation. To address these issues, we propose CR-Walker in this paper, a model that performs tree-structured reasoning on a knowledge graph, and generates informative dialog acts to guide language generation. The unique scheme of tree-structured reasoning views the traversed entity at each hop as part of dialog acts to facilitate language generation, which links how entities are selected and expressed. Automatic and human evaluations show that CR-Walker can arrive at more accurate recommendation, and generate more informative and engaging responses.

DIALKI: Knowledge Identification in Conversational Systems through Dialogue-Document Contextualization

Zeqiu Wu et al. 17:00–17:15

Identifying relevant knowledge to be used in conversational systems that are grounded in long documents is critical to effective response generation. We introduce a knowledge identification model that leverages the document structure to provide dialogue-contextualized passage encodings and better locate knowledge relevant to the conversation. An auxiliary loss captures the history of dialogue-document connections. We demonstrate the effectiveness of our model on two document-grounded conversational datasets and provide analyses showing generalization to unseen documents and long dialogue contexts.

Iconary: A Pictionary-Based Game for Testing Multimodal Communication with Drawings and Text

Christopher Clark et al. 17:15–17:30

Communicating with humans is challenging for AIs because it requires a shared understanding of the world, complex semantics (e.g., metaphors or analogies), and at times multi-modal gestures (e.g., pointing with a finger, or an arrow in a diagram). We investigate these challenges in the context of Iconary, a collaborative game of drawing and guessing based on Pictionary, that poses a novel challenge for the research community. In Iconary, a Guesser tries to identify a phrase that a Drawer is drawing by composing icons, and the Drawer iteratively revises the drawing to help the Guesser in response. This back-and-forth often uses canonical scenes, visual metaphor, or icon compositions to express challenging words, making it an ideal test for mixing language and visual/symbolic communication in AI. We propose models to play Iconary and train them on over 55,000 games between human players. Our models are skillful players and are able to employ world knowledge in language models to play with words unseen during training.

Self-training Improves Pre-training for Few-shot Learning in Task-oriented Dialog Systems

Fei Mi et al. 17:30–17:45

As the labeling cost for different modules in task-oriented dialog (ToD) systems is expensive, a major challenge is to train different modules with the least amount of labeled data. Recently, large-scale pre-trained language models, have shown promising results for few-shot learning in ToD. In this paper, we devise a self-training approach to utilize the abundant unlabeled dialog data to further improve state-of-the-art pre-trained models in few-shot learning scenarios for ToD systems. Specifically, we propose a self-training approach that iteratively labels the most confident unlabeled data to train a stronger Student model. Moreover, a new text augmentation technique (GradAug) is proposed to better train the Student by replacing non-crucial tokens using a masked language model. We conduct extensive experiments and present analyses on four downstream tasks in ToD, including intent classification, dialog state tracking, dialog act prediction, and response selection. Empirical results demonstrate that the proposed self-training approach consistently improves state-of-the-art pre-trained models (BERT, ToD-BERT) when only a small number of labeled data are available.

Contextual Rephrase Detection for Reducing Friction in Dialogue Systems

Zhuoyi Wang et al. 17:45–17:55

For voice assistants like Alexa, Google Assistant, and Siri, correctly interpreting users' intentions is of utmost importance. However, users sometimes experience friction with these assistants, caused by errors from different system components or user errors such as slips of the tongue. Users tend to rephrase their queries until they get a satisfactory response. Rephrase detection is used to identify the rephrases and has long been treated as a task with pairwise input, which does not fully utilize the contextual information (e.g. users' implicit feedback). To this end, we propose a contextual rephrase detection model ContReph to automatically identify rephrases from multi-turn dialogues. We showcase how to leverage the dialogue context and user-agent interaction signals, including the user's implicit feedback and the time gap between different turns, which can help significantly outperform the pairwise rephrase detection models.

Few-Shot Intent Detection via Contrastive Pre-Training and Fine-Tuning
Jianguo Zhang et al. 17:55–18:05

In this work, we focus on a more challenging few-shot intent detection scenario where many intents are fine-grained and semantically similar. We present a simple yet effective few-shot intent detection schema via contrastive pre-training and fine-tuning. Specifically, we first conduct self-supervised contrastive pre-training on collected intent datasets, which implicitly learns to discriminate semantically similar utterances without using any labels. We then perform few-shot intent detection together with supervised contrastive learning, which explicitly pulls utterances from the same intent closer and pushes utterances across different intents farther. Experimental results show that our proposed method achieves state-of-the-art performance on three challenging intent detection datasets under 5-shot and 10-shot settings.

"It doesn't look good for a date": Transforming Critiques into Preferences for Conversational Recommendation Systems
Victor Bursztyn et al. 18:05–18:15

Conversations aimed at determining good recommendations are iterative in nature. People often express their preferences in terms of a critique of the current recommendation (e.g., "It doesn't look good for a date"), requiring some degree of common sense for a preference to be inferred. In this work, we present a method for transforming a user critique into a positive preference (e.g., "I prefer more romantic") in order to retrieve reviews pertaining to potentially better recommendations (e.g., "Perfect for a romantic dinner"). We leverage a large neural language model (LM) in a few-shot setting to perform critique-to-preference transformation, and we test two methods for retrieving recommendations: one that matches embeddings, and another that fine-tunes an LM for the task. We instantiate this approach in the restaurant domain and evaluate it using a new dataset of restaurant critiques. In an ablation study, we show that utilizing critique-to-preference transformation improves recommendations, and that there are at least three general cases that explain this improved performance.

Session 4C: Information Extraction 3

Bavaro 3 *Chair: Manuel R. Ciosici*

AttentionRank: Unsupervised Keyphrase Extraction using Self and Cross Attentions

Haoran Ding and Xiao Luo 16:45–17:00

Keyword or keyphrase extraction is to identify words or phrases presenting the main topics of a document. This paper proposes the AttentionRank, a hybrid attention model, to identify keyphrases from a document in an unsupervised manner. AttentionRank calculates self-attention and cross-attention using a pre-trained language model. The self-attention is designed to determine the importance of a candidate within the context of a sentence. The cross-attention is calculated to identify the semantic relevance between a candidate and sentences within a document. We evaluate the AttentionRank on three publicly available datasets against seven baselines. The results show that the AttentionRank is an effective and robust unsupervised keyphrase extraction model on both long and short documents. Source code is available on Github.

Unsupervised Relation Extraction: A Variational Autoencoder Approach

Chenhan Yuan and Hoda Eldardiry 17:00–17:15

Unsupervised relation extraction works by clustering entity pairs that have the same relations in the text. Some existing variational autoencoder (VAE)-based approaches train the relation extraction model as an encoder that generates relation classifications. A decoder is trained along with the encoder to reconstruct the encoder input based on the encoder-generated relation classifications. These classifications are a latent variable so they are required to follow a pre-defined prior distribution which results in unstable training. We propose a VAE-based unsupervised relation extraction technique that overcomes this limitation by using the classifications as an intermediate variable instead of a latent variable. Specifically, classifications are conditioned on sentence input, while the latent variable is conditioned on both the classifications and the sentence input. This allows our model to connect the decoder with the encoder without putting restrictions on the classification distribution; which improves training stability. Our approach is evaluated on the NYT dataset and outperforms state-of-the-art methods.

Robust Retrieval Augmented Generation for Zero-shot Slot Filling

Michael Glass et al. 17:15–17:30

Automatically inducing high quality knowledge graphs from a given collection of documents still remains a challenging problem in AI. One way to make headway for this problem is through advancements in a related task known as slot filling. In this task, given an entity query in form of [Entity, Slot, ?], a system is asked to 'fill' the slot by generating or extracting the missing value exploiting evidence extracted from relevant passage(s) in the given document collection. The recent works in the field try to solve this task in an end-to-end fashion using retrieval-based language models. In this paper, we present a novel approach to zero-shot slot filling that extends dense passage retrieval with hard negatives and robust training procedures for retrieval augmented generation models. Our model reports large improvements on both T-REx and zsRE slot filling datasets, improving both passage retrieval and slot value generation, and ranking at the top-1 position in the KILT leaderboard. Moreover, we demonstrate the robustness of our system showing its domain adaptation capability on a new variant of the TACRED dataset for slot filling, through a combination of zero/few-shot learning. We release the source code and pre-trained models.

Everything Is All It Takes: A Multipronged Strategy for Zero-Shot Cross-Lingual Information Extraction

Mahsa Yarmohammadi et al. 17:30–17:45

Zero-shot cross-lingual information extraction (IE) describes the construction of an IE model for some target language, given existing annotations exclusively in some other language, typically English. While the advance of pretrained multilingual encoders suggests an easy optimism of "train on English, run on any language", we find through a thorough exploration and extension of techniques that a combination of approaches, both new and old, leads to better performance than any one cross-lingual strategy in particular. We explore techniques including data projection and self-training, and how different pretrained encoders impact them. We use English-to-Arabic IE as our initial example, demonstrating strong performance in this setting for event extraction, named entity recognition, part-of-speech tagging, and dependency parsing. We then apply data projection and self-training to three tasks across eight target languages. Because no single set of techniques performs the best across all tasks, we encourage practitioners to explore various configurations of the techniques described in this work when seeking to improve on zero-shot training.

Partially Supervised Named Entity Recognition via the Expected Entity Ratio Loss

Thomas Effland and Michael Collins 17:45–18:00

We study learning named entity recognizers in the presence of missing entity annotations. We approach this setting as tagging with latent variables and propose a novel loss, the Expected Entity Ratio, to learn models in the presence of systematically missing tags. We show that our approach is both theoretically sound and empirically useful. Experimentally, we find that it outperforms previous state-of-the-art baselines both in accuracy and speed across a variety of languages, annotation scenarios, and amounts of labeled data. In particular, we find that it outperforms the previous state of the art by $+12.7$ F1 score in a challenging setting with only $1,000$ biased annotations. We also show that, when combined with our approach, a novel "quick and sparse" annotation scheme outperforms exhaustive annotation for modest annotation budgets.

MasakhaNER: Named Entity Recognition for African Languages
Jade Abbott et al. 18:00–18:15
We take a step towards addressing the under-representation of the African continent in NLP research, by bringing together different stakeholders to create the first large, publicly available, high-quality dataset for named entity recognition (NER) in ten African languages. We detail the characteristics of these languages to help researchers and practitioners better understand the challenges they pose for NER tasks. We analyze our datasets and conduct an extensive empirical evaluation of state-of-the-art methods across both supervised and transfer learning settings. Finally, we release the data, code, and models to inspire future research on African NLP.

Session 4D: Ethics and NLP

Punta Cana 1,2 *Chair: Sasha Luccioni*

Harms of Gender Exclusivity and Challenges in Non-Binary Representation in Language Technologies

Sunipa Dev et al. 16:45–17:00

Gender is widely discussed in the context of language tasks and when examining the stereotypes propagated by language models. However, current discussions primarily treat gender as binary, which can perpetuate harms such as the cyclical erasure of non-binary gender identities. These harms are driven by model and dataset biases, which are consequences of the non-recognition and lack of understanding of non-binary genders in society. In this paper, we explain the complexity of gender and language around it, and survey non-binary persons to understand harms associated with the treatment of gender as binary in English language technologies. We also detail how current language representations (e.g., GloVe, BERT) capture and perpetuate these harms and related challenges that need to be acknowledged and addressed for representations to equitably encode gender information.

Are Gender-Neutral Queries Really Gender-Neutral? Mitigating Gender Bias in Image Search

Jialu Wang, Yang Liu, and Xin Wang 17:00–17:15

Internet search affects people's cognition of the world, so mitigating biases in search results and learning fair models is imperative for social good. We study a unique gender bias in image search in this work: the search images are often gender-imbalanced for gender-neutral natural language queries. We diagnose two typical image search models, the specialized model trained on in-domain datasets and the generalized representation model pre-trained on massive image and text data across the internet. Both models suffer from severe gender bias. Therefore, we introduce two novel debiasing approaches: an in-processing fair sampling method to address the gender imbalance issue for training models, and a post-processing feature clipping method base on mutual information to debias multimodal representations of pre-trained models. Extensive experiments on MS-COCO and Flickr30K benchmarks show that our methods significantly reduce the gender bias in image search models.

Style Pooling: Automatic Text Style Obfuscation for Improved Classification Fairness

Fatemehsadat Mireshghallah and Taylor Berg-Kirkpatrick 17:15–17:30

Text style can reveal sensitive attributes of the author (e.g. age and race) to the reader, which can, in turn, lead to privacy violations and bias in both human and algorithmic decisions based on text. For example, the style of writing in job applications might reveal protected attributes of the candidate which could lead to bias in hiring decisions, regardless of whether hiring decisions are made algorithmically or by humans. We propose a VAE-based framework that obfuscates stylistic features of human-generated text through style transfer, by automatically re-writing the text itself. Critically, our framework operationalizes the notion of obfuscated style in a flexible way that enables two distinct notions of obfuscated style: (1) a minimal notion that effectively intersects the various styles seen in training, and (2) a maximal notion that seeks to obfuscate by adding stylistic features of all sensitive attributes to text, in effect, computing a union of styles. Our style-obfuscation framework can be used for multiple purposes, however, we demonstrate its effectiveness in improving the fairness of downstream classifiers. We also conduct a comprehensive study on style-pooling's effect on fluency, semantic consistency, and attribute removal from text, in two and three domain style transfer.

Modeling Disclosive Transparency in NLP Application Descriptions

Michael Saxon et al. 17:30–17:45

Broader disclosive transparency—truth and clarity in communication regarding the function of AI systems—is widely considered desirable. Unfortunately, it is a nebulous concept, difficult to both define and quantify. This is problematic, as previous work has demonstrated possible trade-offs and negative consequences to disclosive transparency, such as a confusion effect, where "too much information" clouds a reader's understanding of what a system description means. Disclosive transparency's subjective nature has rendered deep study into these problems and their remedies difficult. To improve this state of affairs, We introduce neural language model-based probabilistic metrics to directly model disclosive transparency, and demonstrate that they correlate with user and expert opinions of system transparency, making them a valid objective proxy. Finally, we demonstrate the use of these metrics in a pilot study quantifying the relationships between transparency, confusion, and user perceptions in a corpus of real NLP system descriptions.

Reconstruction Attack on Instance Encoding for Language Understanding

Shangyu Xie and Yuan Hong 17:45–17:55

A private learning scheme TextHide was recently proposed to protect the private text data during the training phase via so-called instance encoding. We propose a novel reconstruction attack to break TextHide by recovering the private training data, and thus unveil the privacy risks of instance encoding. We have experimentally validated the effectiveness of the reconstruction attack with two commonly-used datasets for sentence classification. Our attack would advance the development of privacy preserving machine learning in the context of natural language processing.

Fairness-aware Class Imbalanced Learning
Shivashankar Subramanian et al. 17:55–18:05

Class imbalance is a common challenge in many NLP tasks, and has clear connections to bias, in that bias in training data often leads to higher accuracy for majority groups at the expense of minority groups. However there has traditionally been a disconnect between research on class-imbalanced learning and mitigating bias, and only recently have the two been looked at through a common lens. In this work we evaluate long-tail learning methods for tweet sentiment and occupation classification, and extend a margin-loss based approach with methods to enforce fairness. We empirically show through controlled experiments that the proposed approaches help mitigate both class imbalance and demographic biases.

CRYPTOGRU: Low Latency Privacy-Preserving Text Analysis With GRU
Bo Feng et al. 18:05–18:15

Homomorphic encryption (HE) and garbled circuit (GC) provide the protection for users' privacy. However, simply mixing the HE and GC in RNN models suffer from long inference latency due to slow activation functions. In this paper, we present a novel hybrid structure of HE and GC gated recurrent unit (GRU) network, , for low-latency secure inferences. replaces computationally expensive GC-based $tanh$ with fast GC-based $ReLU$, and then quantizes $sigmoid$ and $ReLU$ to smaller bit-length to accelerate activations in a GRU. We evaluate with multiple GRU models trained on 4 public datasets. Experimental results show achieves top-notch accuracy and improves the secure inference latency by up to $138\times$ over one of the state-of-the-art secure networks on the Penn Treebank dataset.

Session 4E: Phonology, Morphology and Word Segmentation

Punta Cana 3,4 *Chair: Xinchi Chen*

Local Word Discovery for Interactive Transcription

William Lane and Steven Bird 16:45–17:00

Human expertise and the participation of speech communities are essential factors in the success of technologies for low-resource languages. Accordingly, we propose a new computational task which is tuned to the available knowledge and interests in an Indigenous community, and which supports the construction of high quality texts and lexicons. The task is illustrated for Kunwinjku, a morphologically-complex Australian language. We combine a finite state implementation of a published grammar with a partial lexicon, and apply this to a noisy phone representation of the signal. We locate known lexemes in the signal and use the morphological transducer to build these out into hypothetical, morphologically-complex words for human validation. We show that applying a single iteration of this method results in a relative transcription density gain of 17%. Further, we find that 75% of breath groups in the test set receive at least one correct partial or full-word suggestion.

Segment, Mask, and Predict: Augmenting Chinese Word Segmentation with Self-Supervision

Mieradilijiang Maimaiti et al. 17:00–17:15

Recent state-of-the-art (SOTA) effective neural network methods and fine-tuning methods based on pre-trained models (PTM) have been used in Chinese word segmentation (CWS), and they achieve great results. However, previous works focus on training the models with the fixed corpus at every iteration. The intermediate generated information is also valuable. Besides, the robustness of the previous neural methods is limited by the large-scale annotated data. There are a few noises in the annotated corpus. Limited efforts have been made by previous studies to deal with such problems. In this work, we propose a self-supervised CWS approach with a straightforward and effective architecture. First, we train a word segmentation model and use it to generate the segmentation results. Then, we use a revised masked language model (MLM) to evaluate the quality of the segmentation results based on the predictions of the MLM. Finally, we leverage the evaluations to aid the training of the segmenter by improved minimum risk training. Experimental results show that our approach outperforms previous methods on 9 different CWS datasets with single criterion training and multiple criteria training and achieves better robustness.

Minimal Supervision for Morphological Inflection

Omer Goldman and Reut Tsarfaty 17:15–17:30

Neural models for the various flavours of morphological reinflection tasks have proven to be extremely accurate given ample labeled data, yet labeled data may be slow and costly to obtain. In this work we aim to overcome this annotation bottleneck by bootstrapping labeled data from a seed as small as *five* labeled inflection tables, accompanied by a large bulk of unlabeled text. Our bootstrapping method exploits the orthographic and semantic regularities in morphological systems in a two-phased setup, where word tagging based on *analogies* is followed by word pairing based on *distances*. Our experiments with the Paradigm Cell Filling Problem over eight typologically different languages show that in languages with relatively simple morphology, orthographic regularities on their own allow inflection models to achieve respectable accuracy. Combined orthographic and semantic regularities alleviate difficulties with particularly complex morpho-phonological systems. We further show that our bootstrapping methods substantially outperform hallucination-based methods commonly used for overcoming the annotation bottleneck in morphological reinflection tasks.

Fast WordPiece Tokenization

Xinying Song et al. 17:30–17:45

Tokenization is a fundamental preprocessing step for almost all NLP tasks. In this paper, we propose efficient algorithms for the WordPiece tokenization used in BERT, from single-word tokenization to general text (e.g., sentence) tokenization. When tokenizing a single word, WordPiece uses a longest-match-first strategy, known as maximum matching. The best known algorithms so far are On^2 (where n is the input length) or O(nm) (where m is the maximum vocabulary token length). We propose a novel algorithm whose tokenization complexity is strictly O(n). Our method is inspired by the Aho-Corasick algorithm. We introduce additional linkages on top of the trie built from the vocabulary, allowing smart transitions when the trie matching cannot continue. For general text, we further propose an algorithm that combines pre-tokenization (splitting the text into words) and our linear-time WordPiece method into a single pass. Experimental results show that our method is 8.2x faster than HuggingFace Tokenizers and 5.1x faster than TensorFlow Text on average for general text tokenization.

You should evaluate your language model on marginal likelihood over tokenisations

Kris Cao and Laura Rimell 17:45–18:00

Neural language models typically tokenise input text into sub-word units to achieve an open vocabulary. The standard approach is to use a single canonical tokenisation at both train and test time. We suggest that this approach is unsatisfactory and may bottleneck our evaluation of language model performance. Using only the one-best tokenisation ignores tokeniser uncertainty over alternative tokenisations, which may hurt model out-of-domain performance. In this paper, we argue that instead, language models should be evaluated on their marginal likelihood over tokenisations. We compare different estimators for the marginal likelihood

based on sampling, and show that it is feasible to estimate the marginal likelihood with a manageable number of samples. We then evaluate a pretrained language model on both the one-best-tokenisation and marginal perplexities, and show that the marginal perplexity can be significantly better than the one best, especially on out-of-domain data. We link this difference in perplexity to the tokeniser uncertainty as measured by tokeniser entropy. We discuss some implications of our results for language model training and evaluation, particularly with regard to tokenisation robustness.

Identity-Based Patterns in Deep Convolutional Networks: Generative Adversarial Phonology and Reduplication

Gašper Beguš 18:00–18:15

This paper models unsupervised learning of an identity-based pattern (or copying) in speech called reduplication from raw continuous data with deep convolutional neural networks. We use the ciwGAN architecture Begus (2021a) in which learning of meaningful representations in speech emerges from a requirement that the CNNs generate informative data. We propose a technique to wug-test CNNs trained on speech and, based on four generative tests, argue that the network learns to represent an identity-based pattern in its latent space. By manipulating only two categorical variables in the latent space, we can actively turn an unreduplicated form into a reduplicated form with no other substantial changes to the output in the majority of cases. We also argue that the network extends the identity-based pattern to unobserved data. Exploration of how meaningful representations of identity-based patterns emerge in CNNs and how the latent space variables outside of the training range correlate with identity-based patterns in the output has general implications for neural network interpretability.

Session 4F: Speech, Vision, Robotics, Multimodal Grounding 1

Santo Domingo *Chair: Douwe Kiela*

Broaden the Vision: Geo-Diverse Visual Commonsense Reasoning

Da Yin et al. 16:45–17:00

Commonsense is defined as the knowledge on which everyone agrees. However, certain types of commonsense knowledge are correlated with culture and geographic locations and they are only shared locally. For example, the scenes of wedding ceremonies vary across regions due to different customs influenced by historical and religious factors. Such regional characteristics, however, are generally omitted in prior work. In this paper, we construct a Geo-Diverse Visual Commonsense Reasoning dataset (GD-VCR) to test vision-and-language models' ability to understand cultural and geo-location-specific commonsense. In particular, we study two state-of-the-art Vision-and-Language models, VisualBERT and ViLBERT trained on VCR, a standard benchmark with images primarily from Western regions. We then evaluate how well the trained models can generalize to answering the questions in GD-VCR. We find that the performance of both models for non-Western regions including East Asia, South Asia, and Africa is significantly lower than that for Western region. We analyze the reasons behind the performance disparity and find that the performance gap is larger on QA pairs that: 1) are concerned with culture-related scenarios, e.g., weddings, religious activities, and festivals; 2) require high-level geo-diverse commonsense reasoning rather than low-order perception and recognition. Dataset and code are released at https://github.com/WadeYin9712/GD-VCR.

Reference-Centric Models for Grounded Collaborative Dialogue

Daniel Fried, Justin Chiu, and Dan Klein 17:00–17:15

We present a grounded neural dialogue model that successfully collaborates with people in a partially-observable reference game. We focus on a setting where two agents each observe an overlapping part of a world context and need to identify and agree on some object they share. Therefore, the agents should pool their information and communicate pragmatically to solve the task. Our dialogue agent accurately grounds referents from the partner's utterances using a structured reference resolver, conditions on these referents using a recurrent memory, and uses a pragmatic generation procedure to ensure the partner can resolve the references the agent produces. We evaluate on the OneCommon spatial grounding dialogue task (Udagawa and Aizawa 2019), involving a number of dots arranged on a board with continuously varying positions, sizes, and shades. Our agent substantially outperforms the previous state of the art for the task, obtaining a 20% relative improvement in successful task completion in self-play evaluations and a 50% relative improvement in success in human evaluations.

CrossVQA: Scalably Generating Benchmarks for Systematically Testing VQA Generalization

Arjun Akula et al. 17:15–17:30

One challenge in evaluating visual question answering (VQA) models in the cross-dataset adaptation setting is that the distribution shifts are multi-modal, making it difficult to identify if it is the shifts in visual or language features that play a key role. In this paper, we propose a semi-automatic framework for generating disentangled shifts by introducing a controllable visual question-answer generation (VQAG) module that is capable of generating highly-relevant and diverse question-answer pairs with the desired dataset style. We use it to create CrossVQA, a collection of test splits for assessing VQA generalization based on the VQA2, VizWiz, and Open~Images datasets. We provide an analysis of our generated datasets and demonstrate its utility by using them to evaluate several state-of-the-art VQA systems. One important finding is that the visual shifts in cross-dataset VQA matter more than the language shifts. More broadly, we present a scalable framework for systematically evaluating the machine with little human intervention.

Generative Spoken Language Modeling from Raw Audio

Kushal Lakhotia et al. 17:30–17:45

We introduce Generative Spoken Language Modeling, the task of jointly learning the acoustic and linguistic characteristics of a language from raw audio (without text), and a set of metrics to automatically evaluate the learned representations at acoustic and linguistic levels for both encoding and generation. We set up baseline systems consisting of a discrete speech encoder (returning pseudo-text units), a generative language model (trained on pseudo-text), and a speech decoder (generating a waveform from pseudo-text) and validate the proposed metrics with human evaluation. Across three unsupervised speech encoders (CPC, wav2vec 2.0, HuBERT), we find that the number of discrete units (50, 100, or 200) matters in a task-dependent and encoderdependant way, and that some combinations approach text-based topline systems.

Visual Goal-Step Inference using wikiHow

Yue Yang et al. 17:45–17:55

Understanding what sequence of steps are needed to complete a goal can help artificial intelligence systems reason about human activities. Past work in NLP has examined the task of goal-step inference for text. We introduce the visual analogue. We propose the Visual Goal-Step Inference (VGSI) task, where a model is given a textual goal and must choose which of four images represents a plausible step towards that goal. With a new dataset harvested from wikiHow consisting of 772,277 images representing human actions, we show that our task is challenging for state-of-the-art multimodal models. Moreover, the multimodal representation learned from our data can be effectively transferred to other datasets like HowTo100m, increasing the VGSI accuracy

by 15 - 20%. Our task will facilitate multimodal reasoning about procedural events.

Systematic Generalization on gSCAN: What is Nearly Solved and What is Next?
Linlu Qiu et al. 17:55–18:05

We analyze the *grounded* SCAN (gSCAN) benchmark, which was recently proposed to study systematic generalization for grounded language understanding. First, we study which aspects of the original benchmark can be solved by commonly used methods in multi-modal research. We find that a general-purpose Transformer-based model with cross-modal attention achieves strong performance on a majority of the gSCAN splits, surprisingly outperforming more specialized approaches from prior work. Furthermore, our analysis suggests that many of the remaining errors reveal the same fundamental challenge in systematic generalization of linguistic constructs regardless of visual context. Second, inspired by this finding, we propose challenging new tasks for gSCAN by generating data to incorporate relations between objects in the visual environment. Finally, we find that current models are surprisingly data inefficient given the narrow scope of commands in gSCAN, suggesting another challenge for future work.

Effect of Visual Extensions on Natural Language Understanding in Vision-and-Language Models
Taichi Iki and Akiko Aizawa 18:05–18:15

A method for creating a vision-and-language (V&L) model is to extend a language model through structural modifications and V&L pre-training. Such an extension aims to make a V&L model inherit the capability of natural language understanding (NLU) from the original language model. To see how well this is achieved, we propose to evaluate V&L models using an NLU benchmark (GLUE). We compare five V&L models, including single-stream and dual-stream models, trained with the same pre-training. Dual-stream models, with their higher modality independence achieved by approximately doubling the number of parameters, are expected to preserve the NLU capability better. Our main finding is that the dual-stream scores are not much different than the single-stream scores, contrary to expectation. Further analysis shows that pre-training causes the performance drop in NLU tasks with few exceptions. These results suggest that adopting a single-stream structure and devising the pre-training could be an effective method for improving the maintenance of language knowledge in V&L extensions.

Virtual Poster and Demo Session I

Time: *19:00–21:00*

Neural Path Hunter: Reducing Hallucination in Dialogue Systems via Path Grounding
Nouha Dziri et al.

Dialogue systems powered by large pre-trained language models exhibit an innate ability to deliver fluent and natural-sounding responses. Despite their impressive performance, these models are fitful and can often generate factually incorrect statements impeding their widespread adoption. In this paper, we focus on the task of improving faithfulness and reducing hallucination of neural dialogue systems to known facts supplied by a Knowledge Graph (KG). We propose Neural Path Hunter which follows a generate-then-refine strategy whereby a generated response is amended using the KG. Neural Path Hunter leverages a separate token-level fact critic to identify plausible sources of hallucination followed by a refinement stage that retrieves correct entities by crafting a query signal that is propagated over a k-hop subgraph. We empirically validate our proposed approach on the OpenDialKG dataset (Moon et al., 2019) against a suite of metrics and report a relative improvement of faithfulness over dialogue responses by 20.35% based on FeQA (Durmus et al., 2020). The code is available at https://github.com/nouhadziri/Neural-Path-Hunter.

Thinking Clearly, Talking Fast: Concept-Guided Non-Autoregressive Generation for Open-Domain Dialogue Systems
Yicheng Zou et al.

Human dialogue contains evolving concepts, and speakers naturally associate multiple concepts to compose a response. However, current dialogue models with the seq2seq framework lack the ability to effectively manage concept transitions and can hardly introduce multiple concepts to responses in a sequential decoding manner. To facilitate a controllable and coherent dialogue, in this work, we devise a concept-guided non-autoregressive model (CG-nAR) for open-domain dialogue generation. The proposed model comprises a multi-concept planning module that learns to identify multiple associated concepts from a concept graph and a customized Insertion Transformer that performs concept-guided non-autoregressive generation to complete a response. The experimental results on two public datasets show that CG-nAR can produce diverse and coherent responses, outperforming state-of-the-art baselines in both automatic and human evaluations with substantially faster inference speed.

Perspective-taking and Pragmatics for Generating Empathetic Responses Focused on Emotion Causes
Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim

Empathy is a complex cognitive ability based on the reasoning of others' affective states. In order to better understand others and express stronger empathy in dialogues, we argue that two issues must be tackled at the same time: (i) identifying which word is the cause for the other's emotion from his or her utterance and (ii) reflecting those specific words in the response generation. However, previous approaches for recognizing emotion cause words in text require sub-utterance level annotations, which can be demanding. Taking inspiration from social cognition, we leverage a generative estimator to infer emotion cause words from utterances with no word-level label. Also, we introduce a novel method based on pragmatics to make dialogue models focus on targeted words in the input during generation. Our method is applicable to any dialogue models with no additional training on the fly. We show our approach improves multiple best-performing dialogue agents on generating more focused empathetic responses in terms of both automatic and human evaluation.

Generation and Extraction Combined Dialogue State Tracking with Hierarchical Ontology Integration
Xinmeng Li et al.

Recently, the focus of dialogue state tracking has expanded from single domain to multiple domains. The task is characterized by the shared slots between domains. As the scenario gets more complex, the out-of-vocabulary problem also becomes severer. Current models are not satisfactory for solving the challenges of ontology integration between domains and out-of-vocabulary problems. To address the problem, we explore the hierarchical semantic of ontology and enhance the interrelation between slots with masked hierarchical attention. In state value decoding stage, we solve the out-of-vocabulary problem by combining generation method and extraction method together. We evaluate the performance of our model on two representative datasets, MultiWOZ in English and CrossWOZ in Chinese. The results show that our model yields a significant performance gain over current state-of-the-art state tracking model and it is more robust to out-of-vocabulary problem compared with other methods.

CoLV: A Collaborative Latent Variable Model for Knowledge-Grounded Dialogue Generation
Haolan Zhan et al.

Knowledge-grounded dialogue generation has achieved promising performance with the engagement of external knowledge sources. Typical approaches towards this task usually perform relatively independent two subtasks, i.e., knowledge selection and knowledge-aware response generation. In this paper, in order to improve the diversity of both knowledge selection and knowledge-aware response generation, we propose a collabora-

tive latent variable (CoLV) model to integrate these two aspects simultaneously in separate yet collaborative latent spaces, so as to capture the inherent correlation between knowledge selection and response generation. During generation, our proposed model firstly draws knowledge candidate from the latent space conditioned on the dialogue context, and then samples a response from another collaborative latent space conditioned on both the context and the selected knowledge. Experimental results on two widely-used knowledge-grounded dialogue datasets show that our model outperforms previous methods on both knowledge selection and response generation.

A Three-Stage Learning Framework for Low-Resource Knowledge-Grounded Dialogue Generation
Shilei Liu et al.

Neural conversation models have shown great potentials towards generating fluent and informative responses by introducing external background knowledge. Nevertheless, it is laborious to construct such knowledge-grounded dialogues, and existing models usually perform poorly when transfer to new domains with limited training samples. Therefore, building a knowledge-grounded dialogue system under the low-resource setting is a still crucial issue. In this paper, we propose a novel three-stage learning framework based on weakly supervised learning which benefits from large scale ungrounded dialogues and unstructured knowledge base. To better cooperate with this framework, we devise a variant of Transformer with decoupled decoder which facilitates the disentangled learning of response generation and knowledge incorporation. Evaluation results on two benchmarks indicate that our approach can outperform other state-of-the-art methods with less training data, and even in zero-resource scenario, our approach still performs well.

Intention Reasoning Network for Multi-Domain End-to-end Task-Oriented Dialogue
Zhiyuan Ma et al.

Recent years has witnessed the remarkable success in end-to-end task-oriented dialog system, especially when incorporating external knowledge information. However, the quality of most existing models' generated response is still limited, mainly due to their lack of fine-grained reasoning on deterministic knowledge (w.r.t. conceptual tokens), which makes them difficult to capture the concept shifts and identify user's real intention in cross-task scenarios. To address these issues, we propose a novel intention mechanism to better model deterministic entity knowledge. Based on such a mechanism, we further propose an intention reasoning network (IR-Net), which consists of joint and multi-hop reasoning, to obtain intention-aware representations of conceptual tokens that can be used to capture the concept shifts involved in task-oriented conversations, so as to effectively identify user's intention and generate more accurate responses. Experimental results verify the effectiveness of IR-Net, showing that it achieves the state-of-the-art performance on two representative multi-domain dialog datasets.

More is Better: Enhancing Open-Domain Dialogue Generation via Multi-Source Heterogeneous Knowledge
Sixing Wu et al.

Despite achieving remarkable performance, previous knowledge-enhanced works usually only use a single-source homogeneous knowledge base of limited knowledge coverage. Thus, they often degenerate into traditional methods because not all dialogues can be linked with knowledge entries. This paper proposes a novel dialogue generation model, MSKE-Dialog, to solve this issue with three unique advantages: (1) Rather than only one, MSKE-Dialog can simultaneously leverage multiple heterogeneous knowledge sources (it includes but is not limited to commonsense knowledge facts, text knowledge, infobox knowledge) to improve the knowledge coverage; (2) To avoid the topic conflict among the context and different knowledge sources, we propose a Multi-Reference Selection to better select context/knowledge; (3) We propose a Multi-Reference Generation to generate informative responses by referring to multiple generation references at the same time. Extensive evaluations on a Chinese dataset show the superior performance of this work against various state-of-the-art approaches. To our best knowledge, this work is the first to use the multi-source heterogeneous knowledge in the open-domain knowledge-enhanced dialogue generation.

Domain-Lifelong Learning for Dialogue State Tracking via Knowledge Preservation Networks
Qingbin Liu et al.

Dialogue state tracking (DST), which estimates user goals given a dialogue context, is an essential component of task-oriented dialogue systems. Conventional DST models are usually trained offline, which requires a fixed dataset prepared in advance. This paradigm is often impractical in real-world applications since online dialogue systems usually involve continually emerging new data and domains. Therefore, this paper explores Domain-Lifelong Learning for Dialogue State Tracking (DLL-DST), which aims to continually train a DST model on new data to learn incessantly emerging new domains while avoiding catastrophically forgetting old learned domains. To this end, we propose a novel domain-lifelong learning method, called Knowledge Preservation Networks (KPN), which consists of multi-prototype enhanced retrospection and multi-strategy knowledge distillation, to solve the problems of expression diversity and combinatorial explosion in the DLL-DST task. Experimental results show that KPN effectively alleviates catastrophic forgetting and outperforms previous state-of-the-art lifelong learning methods by 4.25% and 8.27% of whole joint goal accuracy on the MultiWOZ benchmark and the SGD benchmark, respectively.

CSAGN: Conversational Structure Aware Graph Network for Conversational Semantic Role Labeling
Han Wu, Kun Xu, and Linqi Song

Conversational semantic role labeling (CSRL) is believed to be a crucial step towards dialogue understanding. However, it remains a major challenge for existing CSRL parser to handle conversational structural information. In this paper, we present a simple and effective architecture for CSRL which aims to address this problem. Our model is based on a conversational structure aware graph network which explicitly encodes the speaker dependent information. We also propose a multi-task learning method to further improve the model. Experimental results on benchmark datasets show that our model with our proposed training objectives significantly outperforms previous baselines.

Different Strokes for Different Folks: Investigating Appropriate Further Pre-training Approaches for Diverse Dialogue Tasks
Yao Qiu, Jinchao Zhang, and Jie Zhou

Loading models pre-trained on the large-scale corpus in the general domain and fine-tuning them on specific downstream tasks is gradually becoming a paradigm in Natural Language Processing. Previous investigations prove that introducing a further pre-training phase between pre-training and fine-tuning phases to adapt the model on the domain-specific unlabeled data can bring positive effects. However, most of these further pre-training works just keep running the conventional pre-training task, e.g., masked language model, which can be regarded as the domain adaptation to bridge the data distribution gap. After observing diverse downstream tasks, we suggest that different tasks may also need a further pre-training phase with appropriate training tasks to bridge the task formulation gap. To investigate this, we carry out a study for improving multiple task-oriented dialogue downstream tasks through designing various tasks at the further pre-training phase. The experiment shows that different downstream tasks prefer different further pre-training tasks, which have intrinsic correlation and most further pre-training tasks significantly improve certain target tasks rather than all. Our investigation indicates that it is of great importance and effectiveness to design appropriate further pre-training tasks modeling specific information that benefit downstream tasks. Besides, we present multiple constructive empirical conclusions for enhancing task-oriented dialogues.

Knowledge Enhanced Fine-Tuning for Better Handling Unseen Entities in Dialogue Generation
Leyang Cui et al.

Although pre-training models have achieved great success in dialogue generation, their performance drops dramatically when the input contains an entity that does not appear in pre-training and fine-tuning datasets (unseen entity). To address this issue, existing methods leverage an external knowledge base to generate appropriate responses. In real-world practical, the entity may not be included by the knowledge base or suffer from the precision of knowledge retrieval. To deal with this problem, instead of introducing knowledge base as the input, we force the model to learn a better semantic representation by predicting the information in the knowledge base, only based on the input context. Specifically, with the help of a knowledge base, we introduce two auxiliary training objectives: 1) Interpret Masked Word, which conjectures the meaning of the masked entity given the context; 2) Hypernym Generation, which predicts the hypernym of the entity based on the context. Experiment results on two dialogue corpus verify the effectiveness of our methods under both knowledge available and unavailable settings.

An Evaluation Dataset and Strategy for Building Robust Multi-turn Response Selection Model
Kijong Han, Seojin Lee, and Dong-hun Lee

Multi-turn response selection models have recently shown comparable performance to humans in several benchmark datasets. However, in the real environment, these models often have weaknesses, such as making incorrect predictions based heavily on superficial patterns without a comprehensive understanding of the context. For example, these models often give a high score to the wrong response candidate containing several keywords related to the context but using the inconsistent tense. In this study, we analyze the weaknesses of the open-domain Korean Multi-turn response selection models and publish an adversarial dataset to evaluate these weaknesses. We also suggest a strategy to build a robust model in this adversarial environment.

Unsupervised Conversation Disentanglement through Co-Training
Hui Liu, Zhan Shi, and Xiaodan Zhu

Conversation disentanglement aims to separate intermingled messages into detached sessions, which is a fundamental task in understanding multi-party conversations. Existing work on conversation disentanglement relies heavily upon human-annotated datasets, which is expensive to obtain in practice. In this work, we explore training a conversation disentanglement model without referencing any human annotations. Our method is built upon the deep co-training algorithm, which consists of two neural networks: a message-pair classifier and a session classifier. The former is responsible of retrieving local relations between two messages while the latter categorizes a message to a session by capturing context-aware information. Both the two networks are initialized respectively with pseudo data built from the unannotated corpus. During the deep co-training process, we use the session classifier as a reinforcement learning component to learn a session assigning policy by maximizing the local rewards given by the message-pair classifier. For the message-pair classifier, we enrich its training data by retrieving message pairs with high confidence from the disentangled sessions predicted by

the session classifier. Experimental results on the large Movie Dialogue Dataset demonstrate that our proposed approach achieves competitive performance compared to previous supervised methods. Further experiments show that the predicted disentangled conversations can promote the performance on the downstream task of multi-party response selection.

Don't be Contradicted with Anything! CI-ToD: Towards Benchmarking Consistency for Task-oriented Dialogue System
Libo Qin et al.

Consistency Identification has obtained remarkable success on open-domain dialogue, which can be used for preventing inconsistent response generation. However, in contrast to the rapid development in open-domain dialogue, few efforts have been made to the task-oriented dialogue direction. In this paper, we argue that *consistency problem* is more urgent in task-oriented domain. To facilitate the research, we introduce CI-ToD, a novel dataset for **C**onsistency **I**dentification in **T**ask-**o**riented **D**ialog system. In addition, we not only annotate the single label to enable the model to judge whether the system response is contradictory, but also provide more fine-grained labels (i.e., Dialogue History Inconsistency, User Query Inconsistency and Knowledge Base Inconsistency) to encourage model to know what inconsistent sources lead to it. Empirical results show that state-of-the-art methods only achieve 51.3%, which is far behind the human performance of 93.2%, indicating that there is ample room for improving consistency identification ability. Finally, we conduct exhaustive experiments and qualitative analysis to comprehend key challenges and provide guidance for future directions. All datasets and models are publicly available at `https://github.com/yizhen20133868/CI-ToD`.

Transferable Persona-Grounded Dialogues via Grounded Minimal Edits
Chen Henry Wu et al.

Grounded dialogue models generate responses that are grounded on certain concepts. Limited by the distribution of grounded dialogue data, models trained on such data face the *transferability* challenges in terms of the data distribution and the type of grounded concepts. To address the challenges, we propose the *grounded minimal editing* framework, which minimally edits existing responses to be grounded on the given concept. Focusing on personas, we propose Grounded Minimal Editor (GME), which learns to edit by disentangling and recombining persona-related and persona-agnostic parts of the response. To evaluate persona-grounded minimal editing, we present the PERSONAMI-NEDIT dataset, and experimental results show that GME outperforms competitive baselines by a large margin. To evaluate the transferability, we experiment on the test set of BLENDEDSKILLTALK and show that GME can edit dialogue models' responses to largely improve their persona consistency while preserving the use of knowledge and empathy.

EARL: Informative Knowledge-Grounded Conversation Generation with Entity-Agnostic Representation Learning
Hao Zhou et al.

Generating informative and appropriate responses is challenging but important for building human-like dialogue systems. Although various knowledge-grounded conversation models have been proposed, these models have limitations in utilizing knowledge that infrequently occurs in the training data, not to mention integrating unseen knowledge into conversation generation. In this paper, we propose an Entity-Agnostic Representation Learning (EARL) method to introduce knowledge graphs to informative conversation generation. Unlike traditional approaches that parameterize the specific representation for each entity, EARL utilizes the context of conversations and the relational structure of knowledge graphs to learn the category representation for entities, which is generalized to incorporating unseen entities in knowledge graphs into conversation generation. Automatic and manual evaluations demonstrate that our model can generate more informative, coherent, and natural responses than baseline models.

DialogueCSE: Dialogue-based Contrastive Learning of Sentence Embeddings
Che Liu et al.

Learning sentence embeddings from dialogues has drawn increasing attention due to its low annotation cost and high domain adaptability. Conventional approaches employ the siamese-network for this task, which obtains the sentence embeddings through modeling the context-response semantic relevance by applying a feed-forward network on top of the sentence encoders. However, as the semantic textual similarity is commonly measured through the element-wise distance metrics (e.g. cosine and L2 distance), such architecture yields a large gap between training and evaluating. In this paper, we propose DialogueCSE, a dialogue-based contrastive learning approach to tackle this issue. DialogueCSE first introduces a novel matching-guided embedding (MGE) mechanism, which generates a context-aware embedding for each candidate response embedding (i.e. the context-free embedding) according to the guidance of the multi-turn context-response matching matrices. Then it pairs each context-aware embedding with its corresponding context-free embedding and finally minimizes the contrastive loss across all pairs. We evaluate our model on three multi-turn dialogue datasets: the Microsoft Dialogue Corpus, the Jing Dong Dialogue Corpus, and the E-commerce Dialogue Corpus. Evaluation results show that our approach significantly outperforms the baselines across all three datasets in terms of MAP and Spearman's correlation measures, demonstrating its effectiveness. Further quantitative experiments show that our approach achieves better performance when leveraging more dialogue context and remains robust when less training data is provided.

Improving Graph-based Sentence Ordering with Iteratively Predicted Pairwise Orderings
Shaopeng Lai et al.

Dominant sentence ordering models can be classified into pairwise ordering models and set-to-sequence models. However, there is little attempt to combine these two types of models, which inituitively possess complementary advantages. In this paper, we propose a novel sentence ordering framework which introduces two classifiers to make better use of pairwise orderings for graph-based sentence ordering (Yin et al. 2019, 2021). Specially, given an initial sentence-entity graph, we first introduce a graph-based classifier to predict pairwise orderings between linked sentences. Then, in an iterative manner, based on the graph updated by previously predicted high-confident pairwise orderings, another classifier is used to predict the remaining uncertain pairwise orderings. At last, we adapt a GRN-based sentence ordering model (Yin et al. 2019, 2021) on the basis of final graph. Experiments on five commonly-used datasets demonstrate the effectiveness and generality of our model. Particularly, when equipped with BERT (Devlin et al. 2019) and FHDecoder (Yin et al. 2020), our model achieves state-of-the-art performance. Our code is available at https://github.com/DeepLearnXMU/IRSEG.

Not Just Classification: Recognizing Implicit Discourse Relation on Joint Modeling of Classification and Generation
Feng Jiang et al.

Implicit discourse relation recognition (IDRR) is a critical task in discourse analysis. Previous studies only regard it as a classification task and lack an in-depth understanding of the semantics of different relations. Therefore, we first view IDRR as a generation task and further propose a method joint modeling of the classification and generation. Specifically, we propose a joint model, CG-T5, to recognize the relation label and generate the target sentence containing the meaning of relations simultaneously. Furthermore, we design three target sentence forms, including the question form, for the generation model to incorporate prior knowledge. To address the issue that large discourse units are hardly embedded into the target sentence, we also propose a target sentence construction mechanism that automatically extracts core sentences from those large discourse units. Experimental results both on Chinese MCDTB and English PDTB datasets show that our model CG-T5 achieves the best performance against several state-of-the-art systems.

A Language Model-based Generative Classifier for Sentence-level Discourse Parsing
Ying Zhang, Hidetaka Kamigaito, and Manabu Okumura

Discourse segmentation and sentence-level discourse parsing play important roles for various NLP tasks to consider textual coherence. Despite recent achievements in both tasks, there is still room for improvement due to the scarcity of labeled data. To solve the problem, we propose a language model-based generative classifier (LMGC) for using more information from labels by treating the labels as an input while enhancing label representations by embedding descriptions for each label. Moreover, since this enables LMGC to make ready the representations for labels, unseen in the pre-training step, we can effectively use a pre-trained language model in LMGC. Experimental results on the RST-DT dataset show that our LMGC achieved the state-of-the-art F1 score of 96.72 in discourse segmentation. It further achieved the state-of-the-art relation F1 scores of 84.69 with gold EDU boundaries and 81.18 with automatically segmented boundaries, respectively, in sentence-level discourse parsing.

Multimodal Phased Transformer for Sentiment Analysis
Junyan Cheng et al.

Multimodal Transformers achieve superior performance in multimodal learning tasks. However, the quadratic complexity of the self-attention mechanism in Transformers limits their deployment in low-resource devices and makes their inference and training computationally expensive. We propose multimodal Sparse Phased Transformer (SPT) to alleviate the problem of self-attention complexity and memory footprint. SPT uses a sampling function to generate a sparse attention matrix and compress a long sequence to a shorter sequence of hidden states. SPT concurrently captures interactions between the hidden states of different modalities at every layer. To further improve the efficiency of our method, we use Layer-wise parameter sharing and Factorized Co-Attention that share parameters between Cross Attention Blocks, with minimal impact on task performance. We evaluate our model with three sentiment analysis datasets and achieve comparable or superior performance compared with the existing methods, with a 90% reduction in the number of parameters. We conclude that (SPT) along with parameter sharing can capture multimodal interactions with reduced model size and improved sample efficiency.

Hierarchical Multi-label Text Classification with Horizontal and Vertical Category Correlations
Linli Xu et al.

Hierarchical multi-label text classification (HMTC) deals with the challenging task where an instance can be assigned to multiple hierarchically structured categories at the same time. The majority of prior studies either focus on reducing the HMTC task into a flat multi-label problem ignoring the vertical category correlations or exploiting the dependencies across different hierarchical levels without considering the horizontal correlations among categories at the same level, which inevitably leads to fundamental information loss. In this paper, we propose a novel HMTC framework that considers both vertical and horizontal category correlations. Specifically, we first design a loosely coupled graph convolutional neural network as the representation extractor to obtain representations for words, documents, and, more importantly, level-wise representations for categories, which are not considered in previous works. Then, the learned category representations are adopted to capture the vertical dependencies among levels of category hierarchy and model the horizontal correlations. Finally, based on the document embeddings and category embeddings, we design a hybrid algorithm to predict the categories of the entire hierarchical structure. Extensive experiments conducted on real-world HMTC datasets

validate the effectiveness of the proposed framework with significant improvements over the baselines.

RankNAS: Efficient Neural Architecture Search by Pairwise Ranking
Chi Hu et al.

This paper addresses the efficiency challenge of Neural Architecture Search (NAS) by formulating the task as a ranking problem. Previous methods require numerous training examples to estimate the accurate performance of architectures, although the actual goal is to find the distinction between "good" and "bad" candidates. Here we do not resort to performance predictors. Instead, we propose a performance ranking method (RankNAS) via pairwise ranking. It enables efficient architecture search using much fewer training examples. Moreover, we develop an architecture selection method to prune the search space and concentrate on more promising candidates. Extensive experiments on machine translation and language modeling tasks show that RankNAS can design high-performance architectures while being orders of magnitude faster than state-of-the-art NAS systems.

FLiText: A Faster and Lighter Semi-Supervised Text Classification with Convolution Networks
Chen Liu et al.

In natural language processing (NLP), state-of-the-art (SOTA) semi-supervised learning (SSL) frameworks have shown great performance on deep pre-trained language models such as BERT, and are expected to significantly reduce the demand for manual labeling. However, our empirical studies indicate that these frameworks are not suitable for lightweight models such as TextCNN, LSTM and etc. In this work, we develop a new SSL framework called FLiText, which stands for Faster and Lighter semi-supervised Text classification. FLiText introduces an inspirer network together with the consistency regularization framework, which leverages a generalized regular constraint on the lightweight models for efficient SSL. As a result, FLiText obtains new SOTA performance for lightweight models across multiple SSL benchmarks on text classification. Compared with existing SOTA SSL methods on TextCNN, FLiText improves the accuracy of lightweight model TextCNN from 51.00% to 90.49% on IMDb, 39.8% to 58.06% on Yelp-5, and from 55.3% to 65.08% on Yahoo! Answer. In addition, compared with the fully supervised method on the full dataset, FLiText just uses less than 1% of labeled data to improve the accuracy by 6.59%, 3.94%, and 3.22% on the datasets of IMDb, Yelp-5, and Yahoo! Answer respectively.

Evaluating Debiasing Techniques for Intersectional Biases
Shivashankar Subramanian et al.

Bias is pervasive for NLP models, motivating the development of automatic debiasing techniques. Evaluation of NLP debiasing methods has largely been limited to binary attributes in isolation, e.g., debiasing with respect to binary gender or race, however many corpora involve multiple such attributes, possibly with higher cardinality. In this paper we argue that a truly fair model must consider 'gerrymandering' groups which comprise not only single attributes, but also intersectional groups. We evaluate a form of bias-constrained model which is new to NLP, as well an extension of the iterative nullspace projection technique which can handle multiple identities.

Definition Modelling for Appropriate Specificity
Han Huang, Tomoyuki Kajiwara, and Yuki Arase

Definition generation techniques aim to generate a definition of a target word or phrase given a context. In previous studies, researchers have faced various issues such as the out-of-vocabulary problem and over/under-specificity problems. Over-specific definitions present narrow word meanings, whereas under-specific definitions present general and context-insensitive meanings. Herein, we propose a method for definition generation with appropriate specificity. The proposed method addresses the aforementioned problems by leveraging a pre-trained encoder-decoder model, namely Text-to-Text Transfer Transformer, and introducing a re-ranking mechanism to model specificity in definitions. Experimental results on standard evaluation datasets indicate that our method significantly outperforms the previous state-of-the-art method. Moreover, manual evaluation confirms that our method effectively addresses the over/under-specificity problems.

Transductive Learning for Unsupervised Text Style Transfer
Fei Xiao et al.

Unsupervised style transfer models are mainly based on an inductive learning approach, which represents the style as embeddings, decoder parameters, or discriminator parameters and directly applies these general rules to the test cases. However, the lacking of parallel corpus hinders the ability of these inductive learning methods on this task. As a result, it is likely to cause severe inconsistent style expressions, like 'the salad is rude'. To tackle this problem, we propose a novel transductive learning approach in this paper, based on a retrieval-based context-aware style representation. Specifically, an attentional encoder-decoder with a retriever framework is utilized. It involves top-K relevant sentences in the target style in the transfer process. In this way, we can learn a context-aware style embedding to alleviate the above inconsistency problem. In this paper, both sparse (BM25) and dense retrieval functions (MIPS) are used, and two objective functions are designed to facilitate joint learning. Experimental results show that our method outperforms several strong baselines. The proposed transductive learning approach is general and effective to the task of unsupervised style transfer, and we will apply it to the other two typical methods in the future.

Integrating Semantic Scenario and Word Relations for Abstractive Sentence Summarization
Yong Guan et al.

Recently graph-based methods have been adopted for Abstractive Text Summarization. However, existing graph-based methods only consider either word relations or structure information, which neglect the correlation between them. To simultaneously capture the word relations and structure information from sentences, we propose a novel Dual Graph network for Abstractive Sentence Summarization. Specifically, we first construct semantic scenario graph and semantic word relation graph based on FrameNet, and subsequently learn their representations and design graph fusion method to enhance their correlation and obtain better semantic representation for summary generation. Experimental results show our model outperforms existing state-of-the-art methods on two popular benchmark datasets, i.e., Gigaword and DUC 2004.

Coupling Context Modeling with Zero Pronoun Recovering for Document-Level Natural Language Generation
Xin Tan, Longyin Zhang, and Guodong Zhou

Natural language generation (NLG) tasks on pro-drop languages are known to suffer from zero pronoun (ZP) problems, and the problems remain challenging due to the scarcity of ZP-annotated NLG corpora. In this case, we propose a highly adaptive two-stage approach to couple context modeling with ZP recovering to mitigate the ZP problem in NLG tasks. Notably, we frame the recovery process in a task-supervised fashion where the ZP representation recovering capability is learned during the NLG task learning process, thus our method does not require NLG corpora annotated with ZPs. For system enhancement, we learn an adversarial bot to adjust our model outputs to alleviate the error propagation caused by mis-recovered ZPs. Experiments on three document-level NLG tasks, i.e., machine translation, question answering, and summarization, show that our approach can improve the performance to a great extent, and the improvement on pronoun translation is very impressive.

Adaptive Bridge between Training and Inference for Dialogue Generation
Haoran Xu et al.

Although exposure bias has been widely studied in some NLP tasks, it faces its unique challenges in dialogue response generation, the representative one-to-various generation scenario.In real human dialogue, there are many appropriate responses for the same context, not only with different expressions, but also with different topics. Therefore, due to the much bigger gap between various ground-truth responses and the generated synthetic response, exposure bias is more challenging in dialogue generation task.What's more, as MLE encourages the model to only learn the common words among different ground-truth responses, but ignores the interesting and specific parts, exposure bias may further lead to the common response generation problem, such as "I don't know" and "HaHa?" In this paper, we propose a novel adaptive switching mechanism, which learns to automatically transit between ground-truth learning and generated learning regarding the word-level matching score, such as the cosine similarity. Experimental results on both Chinese STC dataset and English Reddit dataset, show that our adaptive method achieves a significant improvement in terms of metric-based evaluation and human evaluation, as compared with the state-of-the-art exposure bias approaches. Further analysis on NMT task also shows that our model can achieve a significant improvement.

ConRPG: Paraphrase Generation using Contexts as Regularizer
Yuxian Meng et al.

A long-standing issue with paraphrase generation is the lack of reliable supervision signals. In this paper, we propose a new unsupervised paradigm for paraphrase generation based on the assumption that the probabilities of generating two sentences with the same meaning given the same context should be the same. Inspired by this fundamental idea, we propose a pipelined system which consists of paraphrase candidate generation based on contextual language models, candidate filtering using scoring functions, and paraphrase model training based on the selected candidates. The proposed paradigm offers merits over existing paraphrase generation methods: (1) using the context regularizer on meanings, the model is able to generate massive amounts of high-quality paraphrase pairs; (2) the combination of the huge amount of paraphrase candidates and further diversity-promoting filtering yields paraphrases with more lexical and syntactic diversity; and (3) using human-interpretable scoring functions to select paraphrase pairs from candidates, the proposed framework provides a channel for developers to intervene with the data generation process, leading to a more controllable model. Experimental results across different tasks and datasets demonstrate that the proposed paradigm significantly outperforms existing paraphrase approaches in both supervised and unsupervised setups.

Building the Directed Semantic Graph for Coherent Long Text Generation
Ziao Wang, Xiaofeng Zhang, and Hongwei Du

Generating long text conditionally depending on the short input text has recently attracted more and more research efforts. Most existing approaches focus more on introducing extra knowledge to supplement the short input text, but ignore the coherence issue of the generated texts. To address aforementioned research issue, this paper proposes a novel two-stage approach to generate coherent long text. Particularly, we first build a document-level path for each output text with each sentence embedding as its node, and a revised self-organising map (SOM) is proposed to cluster similar nodes of a family of document-level paths to construct the directed semantic graph. Then, three subgraph alignment methods are proposed to extract the maximum matching paths or subgraphs. These directed subgraphs are considered to well preserve extra but relevant content to the short input text, and then they are decoded by the employed pre-trained model to generate

coherent long text. Extensive experiments have been performed on three real-world datasets, and the promising results demonstrate that the proposed approach is superior to the state-of-the-art approaches w.r.t. a number of evaluation criteria.

Iterative GNN-based Decoder for Question Generation

Zichu Fei, Qi Zhang, and Yaqian Zhou

Natural question generation (QG) aims to generate questions from a passage, and generated questions are answered from the passage. Most models with state-of-the-art performance model the previously generated text at each decoding step. However, (1) they ignore the rich structure information that is hidden in the previously generated text. (2) they ignore the impact of copied words on the passage. We perceive that information in previously generated words serves as auxiliary information in subsequent generation. To address these problems, we design the Iterative Graph Network-based Decoder (IGND) to model the previous generation using a Graph Neural Network at each decoding step. Moreover, our graph model captures dependency relations in the passage that boost the generation. Experimental results demonstrate that our model outperforms the state-of-the-art models with sentence-level QG tasks on SQuAD and MARCO datasets.

Asking Questions Like Educational Experts: Automatically Generating Question-Answer Pairs on Real-World Examination Data

Fanyi Qu, Xin Jia, and Yunfang Wu

Generating high quality question-answer pairs is a hard but meaningful task. Although previous works have achieved great results on answer-aware question generation, it is difficult to apply them into practical application in the education field. This paper for the first time addresses the question-answer pair generation task on the real-world examination data, and proposes a new unified framework on RACE. To capture the important information of the input passage we first automatically generate (rather than extracting) keyphrases, thus this task is reduced to keyphrase-question-answer triplet joint generation. Accordingly, we propose a multi-agent communication model to generate and optimize the question and keyphrases iteratively, and then apply the generated question and keyphrases to guide the generation of answers. To establish a solid benchmark, we build our model on the strong generative pre-training model. Experimental results show that our model makes great breakthroughs in the question-answer pair generation task. Moreover, we make a comprehensive analysis on our model, suggesting new directions for this challenging task.

Syntactically-Informed Unsupervised Paraphrasing with Non-Parallel Data

Erguang Yang et al.

Previous works on syntactically controlled paraphrase generation heavily rely on large-scale parallel paraphrase data that is not easily available for many languages and domains. In this paper, we take this research direction to the extreme and investigate whether it is possible to learn syntactically controlled paraphrase generation with nonparallel data. We propose a syntactically-informed unsupervised paraphrasing model based on conditional variational auto-encoder (VAE) which can generate texts in a specified syntactic structure. Particularly, we design a two-stage learning method to effectively train the model using non-parallel data. The conditional VAE is trained to reconstruct the input sentence according to the given input and its syntactic structure. Furthermore, to improve the syntactic controllability and semantic consistency of the pre-trained conditional VAE, we fine-tune it using syntax controlling and cycle reconstruction learning objectives, and employ Gumbel-Softmax to combine these new learning objectives. Experiment results demonstrate that the proposed model trained only on non-parallel data is capable of generating diverse paraphrases with specified syntactic structure. Additionally, we validate the effectiveness of our method for generating syntactically adversarial examples on the sentiment analysis task.

Exploring Task Difficulty for Few-Shot Relation Extraction

Jiale Han, Bo Cheng, and Wei Lu

Few-shot relation extraction (FSRE) focuses on recognizing novel relations by learning with merely a handful of annotated instances. Meta-learning has been widely adopted for such a task, which trains on randomly generated few-shot tasks to learn generic data representations. Despite impressive results achieved, existing models still perform suboptimally when handling hard FSRE tasks, where the relations are fine-grained and similar to each other. We argue this is largely because existing models do not distinguish hard tasks from easy ones in the learning process. In this paper, we introduce a novel approach based on contrastive learning that learns better representations by exploiting relation label information. We further design a method that allows the model to adaptively learn how to focus on hard tasks. Experiments on two standard datasets demonstrate the effectiveness of our method.

MuVER: Improving First-Stage Entity Retrieval with Multi-View Entity Representations

Xinyin Ma et al.

Entity retrieval, which aims at disambiguating mentions to canonical entities from massive KBs, is essential for many tasks in natural language processing. Recent progress in entity retrieval shows that the dual-encoder structure is a powerful and efficient framework to nominate candidates if entities are only identified by descriptions. However, they ignore the property that meanings of entity mentions diverge in different contexts and are related to various portions of descriptions, which are treated equally in previous works. In this work, we propose Multi-View Entity Representations (MuVER), a novel approach for entity retrieval that constructs multi-view representations for entity descriptions and approximates the optimal view for mentions via a heuris-

tic searching method. Our method achieves the state-of-the-art performance on ZESHEL and improves the quality of candidates on three standard Entity Linking datasets.

Treasures Outside Contexts: Improving Event Detection via Global Statistics
Rui Li et al.

Event detection (ED) aims at identifying event instances of specified types in given texts, which has been formalized as a sequence labeling task. As far as we know, existing neural-based ED models make decisions relying entirely on the contextual semantic features of each word in the inputted text, which we find is easy to be confused by the varied contexts in the test stage. To this end, we come up with the idea of introducing a set of statistical features from word-event co-occurrence frequencies in the entire training set to cooperate with contextual features. Specifically, we propose a Semantic and Statistic-Joint Discriminative Network (SS-JDN) consisting of a semantic feature extractor, a statistical feature extractor, and a joint event discriminator. In experiments, SS-JDN effectively exceeds ten recent strong baselines on ACE2005 and KBP2015 datasets. Further, we perform extensive experiments to comprehensively probe SS-JDN.

Uncertain Local-to-Global Networks for Document-Level Event Factuality Identification
Pengfei Cao et al.

Event factuality indicates the degree of certainty about whether an event occurs in the real world. Existing studies mainly focus on identifying event factuality at sentence level, which easily leads to conflicts between different mentions of the same event. To this end, we study the problem of document-level event factuality identification, which determines the event factuality from the view of a document. For this task, we need to consider two important characteristics: Local Uncertainty and Global Structure, which can be utilized to improve performance. In this paper, we propose an Uncertain Local-to-Global Network (ULGN) to make use of these two characteristics. Specifically, we devise a Local Uncertainty Estimation module to model the uncertainty of local information. Moreover, we propose an Uncertain Information Aggregation module to leverage the global structure for integrating the local information. Experimental results demonstrate the effectiveness of our proposed method, outperforming the previous state-of-the-art model by 8.4% and 11.45% of F1 score on two widely used datasets.

A Novel Global Feature-Oriented Relational Triple Extraction Model based on Table Filling
Feiliang Ren et al.

Table filling based relational triple extraction methods are attracting growing research interests due to their promising performance and their abilities on extracting triples from complex sentences. However, this kind of methods are far from their full potential because most of them only focus on using local features but ignore the global associations of relations and of token pairs, which increases the possibility of overlooking some important information during triple extraction. To overcome this deficiency, we propose a global feature-oriented triple extraction model that makes full use of the mentioned two kinds of global associations. Specifically, we first generate a table feature for each relation. Then two kinds of global associations are mined from the generated table features. Next, the mined global associations are integrated into the table feature of each relation. This "generate-mine-integrate" process is performed multiple times so that the table feature of each relation is refined step by step. Finally, each relation's table is filled based on its refined table feature, and all triples linked to this relation are extracted based on its filled table. We evaluate the proposed model on three benchmark datasets. Experimental results show our model is effective and it achieves state-of-the-art results on all of these datasets. The source code of our work is available at: https://github.com/neukg/GRTE.

Structure-Augmented Keyphrase Generation
Jihyuk Kim et al.

This paper studies the keyphrase generation (KG) task for scenarios where structure plays an important role. For example, a scientific publication consists of a short title and a long body, where the title can be used for de-emphasizing unimportant details in the body. Similarly, for short social media posts (, tweets), scarce context can be augmented from titles, though often missing. Our contribution is generating/augmenting structure then injecting these information in the encoding, using existing keyphrases of other documents, complementing missing/incomplete titles. We propose novel structure-augmented document encoding approaches that consist of the following two phases: The first phase, generating structure, extends the given document with related but absent keyphrases, augmenting missing context. The second phase, encoding structure, builds a graph of keyphrases and the given document to obtain the structure-aware representation of the augmented text. Our empirical results validate that our proposed structure augmentation and augmentation-aware encoding/decoding can improve KG for both scenarios, outperforming the state-of-the-art.

An Empirical Study on Multiple Information Sources for Zero-Shot Fine-Grained Entity Typing
Yi Chen et al.

Auxiliary information from multiple sources has been demonstrated to be effective in zero-shot fine-grained entity typing (ZFET). However, there lacks a comprehensive understanding about how to make better use of the existing information sources and how they affect the performance of ZFET. In this paper, we empirically study three kinds of auxiliary information: context consistency, type hierarchy and background knowledge (e.g., prototypes and descriptions) of types, and propose a multi-source fusion model (MSF) targeting these sources.

The performance obtains up to 11.42% and 22.84% absolute gains over state-of-the-art baselines on BBN and Wiki respectively with regard to macro F1 scores. More importantly, we further discuss the characteristics, merits and demerits of each information source and provide an intuitive understanding of the complementarity among them.

DyLex: Incoporating Dynamic Lexicons into BERT for Sequence Labeling

Baojun Wang et al.

Incorporating lexical knowledge into deep learning models has been proved to be very effective for sequence labeling tasks. However, previous works commonly have difficulty dealing with large-scale dynamic lexicons which often cause excessive matching noise and problems of frequent updates. In this paper, we propose DyLex, a plug-in lexicon incorporation approach for BERT based sequence labeling tasks. Instead of leveraging embeddings of words in the lexicon as in conventional methods, we adopt word-agnostic tag embeddings to avoid re-training the representation while updating the lexicon. Moreover, we employ an effective supervised lexical knowledge denoising method to smooth out matching noise. Finally, we introduce a col-wise attention based knowledge fusion mechanism to guarantee the pluggability of the proposed framework. Experiments on ten datasets of three tasks show that the proposed framework achieves new SOTA, even with very large scale lexicons.

MapRE: An Effective Semantic Mapping Approach for Low-resource Relation Extraction

Manqing Dong, Chunguang Pan, and Zhipeng Luo

Neural relation extraction models have shown promising results in recent years; however, the model performance drops dramatically given only a few training samples. Recent works try leveraging the advance in few-shot learning to solve the low resource problem, where they train label-agnostic models to directly compare the semantic similarities among context sentences in the embedding space. However, the label-aware information, i.e., the relation label that contains the semantic knowledge of the relation itself, is often neglected for prediction. In this work, we propose a framework considering both label-agnostic and label-aware semantic mapping information for low resource relation extraction. We show that incorporating the above two types of mapping information in both pretraining and fine-tuning can significantly improve the model performance on low-resource relation extraction tasks.

Heterogeneous Graph Neural Networks for Keyphrase Generation

Jiacheng Ye et al.

The encoder—decoder framework achieves state-of-the-art results in keyphrase generation (KG) tasks by predicting both present keyphrases that appear in the source document and absent keyphrases that do not. However, relying solely on the source document can result in generating uncontrollable and inaccurate absent keyphrases. To address these problems, we propose a novel graph-based method that can capture explicit knowledge from related references. Our model first retrieves some document-keyphrases pairs similar to the source document from a pre-defined index as references. Then a heterogeneous graph is constructed to capture relations with different levels of granularity of the source document and its retrieved references. To guide the decoding process, a hierarchical attention and copy mechanism is introduced, which directly copies appropriate words from both source document and its references based on their relevance and significance. The experimental results on multiple KG benchmarks show that the proposed model achieves significant improvements against other baseline models, especially with regard to the absent keyphrase prediction.

Machine Reading Comprehension as Data Augmentation: A Case Study on Implicit Event Argument Extraction

Jian Liu, Yufeng Chen, and Jinan Xu

Implicit event argument extraction (EAE) is a crucial document-level information extraction task that aims to identify event arguments beyond the sentence level. Despite many efforts for this task, the lack of enough training data has long impeded the study. In this paper, we take a new perspective to address the data sparsity issue faced by implicit EAE, by bridging the task with machine reading comprehension (MRC). Particularly, we devise two data augmentation regimes via MRC, including: 1) implicit knowledge transfer, which enables knowledge transfer from other tasks, by building a unified training framework in the MRC formulation, and 2) explicit data augmentation, which can explicitly generate new training examples, by treating MRC models as an annotator. The extensive experiments have justified the effectiveness of our approach — it not only obtains state-of-the-art performance on two benchmarks, but also demonstrates superior results in a data-low scenario.

Importance Estimation from Multiple Perspectives for Keyphrase Extraction

Mingyang Song, Liping Jing, and Lin Xiao

Keyphrase extraction is a fundamental task in Natural Language Processing, which usually contains two main parts: candidate keyphrase extraction and keyphrase importance estimation. From the view of human understanding documents, we typically measure the importance of phrase according to its syntactic accuracy, information saliency, and concept consistency simultaneously. However, most existing keyphrase extraction approaches only focus on the part of them, which leads to biased results. In this paper, we propose a new approach to estimate the importance of keyphrase from multiple perspectives (called as *KIEMP*) and further improve the performance of keyphrase extraction. Specifically, *KIEMP* estimates the importance of phrase with three modules: a chunking module to measure its syntactic accuracy, a ranking module to check its information saliency, and a matching module to judge the concept (i.e., topic) consistency between phrase and the

whole document. These three modules are seamlessly jointed together via an end-to-end multi-task learning model, which is helpful for three parts to enhance each other and balance the effects of three perspectives. Experimental results on six benchmark datasets show that *KIEMP* outperforms the existing state-of-the-art keyphrase extraction approaches in most cases.

Gradient Imitation Reinforcement Learning for Low Resource Relation Extraction
Xuming Hu et al.

Low-resource Relation Extraction (LRE) aims to extract relation facts from limited labeled corpora when human annotation is scarce. Existing works either utilize self-training scheme to generate pseudo labels that will cause the gradual drift problem, or leverage meta-learning scheme which does not solicit feedback explicitly. To alleviate selection bias due to the lack of feedback loops in existing LRE learning paradigms, we developed a Gradient Imitation Reinforcement Learning method to encourage pseudo label data to imitate the gradient descent direction on labeled data and bootstrap its optimization capability through trial and error. We also propose a framework called GradLRE, which handles two major scenarios in low-resource relation extraction. Besides the scenario where unlabeled data is sufficient, GradLRE handles the situation where no unlabeled data is available, by exploiting a contextualized augmentation method to generate data. Experimental results on two public datasets demonstrate the effectiveness of GradLRE on low resource relation extraction when comparing with baselines.

Low-resource Taxonomy Enrichment with Pretrained Language Models
Kunihiro Takeoka, Kosuke Akimoto, and Masafumi Oyamada

Taxonomies are symbolic representations of hierarchical relationships between terms or entities. While taxonomies are useful in broad applications, manually updating or maintaining them is labor-intensive and difficult to scale in practice. Conventional supervised methods for this enrichment task fail to find optimal parents of new terms in low-resource settings where only small taxonomies are available because of overfitting to hierarchical relationships in the taxonomies. To tackle the problem of low-resource taxonomy enrichment, we propose Musubu, an efficient framework for taxonomy enrichment in low-resource settings with pretrained language models (LMs) as knowledge bases to compensate for the shortage of information. Musubu leverages an LM-based classifier to determine whether or not inputted term pairs have hierarchical relationships. Musubu also utilizes Hearst patterns to generate queries to leverage implicit knowledge from the LM efficiently for more accurate prediction. We empirically demonstrate the effectiveness of our method in extensive experiments on taxonomies from both a SemEval task and real-world retailer datasets.

Entity Relation Extraction as Dependency Parsing in Visually Rich Documents
Yue Zhang et al.

Previous works on key information extraction from visually rich documents (VRDs) mainly focus on labeling the text within each bounding box (i.e.,semantic entity), while the relations in-between are largely unexplored. In this paper, we adapt the popular dependency parsing model, the biaffine parser, to this entity relation extraction task. Being different from the original dependency parsing model which recognizes dependency relations between words, we identify relations between groups of words with layout information instead. We have compared different representations of the semantic entity, different VRD encoders, and different relation decoders. For the model training, we explore multi-task learning to combine entity labeling and relation extraction tasks; and for the evaluation, we conduct experiments on different datasets with filtering and augmentation. The results demonstrate that our proposed model achieves 65.96% F1 score on the FUNSD dataset. As for the real-world application, our model has been applied to the in-house customs data, achieving reliable performance in the production setting.

Synchronous Dual Network with Cross-Type Attention for Joint Entity and Relation Extraction
Hui Wu

Joint entity and relation extraction is challenging due to the complex interaction of interaction between named entity recognition and relation extraction. Although most existing works tend to jointly train these two tasks through a shared network, they fail to fully utilize the interdependence between entity types and relation types. In this paper, we design a novel synchronous dual network (SDN) with cross-type attention via separately and interactively considering the entity types and relation types. On the one hand, SDN adopts two isomorphic bi-directional type-attention LSTM to encode the entity type enhanced representations and the relation type enhanced representations, respectively. On the other hand, SDN explicitly models the interdependence between entity types and relation types via cross-type attention mechanism. In addition, we also propose a new multi-task learning strategy via modeling the interaction of two types of information. Experiments on NYT and WebNLG datasets verify the effectiveness of the proposed model, achieving state-of-the-art performance.

Less is More: Pretrain a Strong Siamese Encoder for Dense Text Retrieval Using a Weak Decoder
Shuqi Lu et al.

Dense retrieval requires high-quality text sequence embeddings to support effective search in the representation space. Autoencoder-based language models are appealing in dense retrieval as they train the encoder to output high-quality embedding that can reconstruct the input texts. However, in this paper, we provide theoretical analyses and show empirically that an autoencoder language model with a low reconstruction loss may not

provide good sequence representations because the decoder may take shortcuts by exploiting language patterns. To address this, we propose a new self-learning method that pre-trains the autoencoder using a *weak* decoder, with restricted capacity and attention flexibility to push the encoder to provide better text representations. Our experiments on web search, news recommendation, and open domain question answering show that our pre-trained model significantly boosts the effectiveness and few-shot ability of dense retrieval models. Our code is available at https://github.com/microsoft/SEED-Encoder/.

TransPrompt: Towards an Automatic Transferable Prompting Framework for Few-shot Text Classification
Chengyu Wang et al.

Recent studies have shown that prompts improve the performance of large pre-trained language models for few-shot text classification. Yet, it is unclear how the prompting knowledge can be transferred across similar NLP tasks for the purpose of mutual reinforcement. Based on continuous prompt embeddings, we propose TransPrompt, a transferable prompting framework for few-shot learning across similar tasks. In TransPrompt, we employ a multi-task meta-knowledge acquisition procedure to train a meta-learner that captures cross-task transferable knowledge. Two de-biasing techniques are further designed to make it more task-agnostic and unbiased towards any tasks. After that, the meta-learner can be adapted to target tasks with high accuracy. Extensive experiments show that TransPrompt outperforms single-task and cross-task strong baselines over multiple NLP tasks and datasets. We further show that the meta-learner can effectively improve the performance on previously unseen tasks; and TransPrompt also outperforms strong fine-tuning baselines when learning with full training sets.

Weakly-supervised Text Classification Based on Keyword Graph
Lu Zhang et al.

Weakly-supervised text classification has received much attention in recent years for it can alleviate the heavy burden of annotating massive data. Among them, keyword-driven methods are the mainstream where user-provided keywords are exploited to generate pseudo-labels for unlabeled texts. However, existing methods treat keywords independently, thus ignore the correlation among them, which should be useful if properly exploited. In this paper, we propose a novel framework called ClassKG to explore keyword-keyword correlation on keyword graph by GNN. Our framework is an iterative process. In each iteration, we first construct a keyword graph, so the task of assigning pseudo labels is transformed to annotating keyword subgraphs. To improve the annotation quality, we introduce a self-supervised task to pretrain a subgraph annotator, and then finetune it. With the pseudo labels generated by the subgraph annotator, we then train a text classifier to classify the unlabeled texts. Finally, we re-extract keywords from the classified texts. Extensive experiments on both long-text and short-text datasets show that our method substantially outperforms the existing ones.

Efficient-FedRec: Efficient Federated Learning Framework for Privacy-Preserving News Recommendation
Jingwei Yi et al.

News recommendation is critical for personalized news access. Most existing news recommendation methods rely on centralized storage of users' historical news click behavior data, which may lead to privacy concerns and hazards. Federated Learning is a privacy-preserving framework for multiple clients to collaboratively train models without sharing their private data. However, the computation and communication cost of directly learning many existing news recommendation models in a federated way are unacceptable for user clients. In this paper, we propose an efficient federated learning framework for privacy-preserving news recommendation. Instead of training and communicating the whole model, we decompose the news recommendation model into a large news model maintained in the server and a light-weight user model shared on both server and clients, where news representations and user model are communicated between server and clients. More specifically, the clients request the user model and news representations from the server, and send their locally computed gradients to the server for aggregation. The server updates its global user model with the aggregated gradients, and further updates its news model to infer updated news representations. Since the local gradients may contain private information, we propose a secure aggregation method to aggregate gradients in a privacy-preserving way. Experiments on two real-world datasets show that our method can reduce the computation and communication cost on clients while keep promising model performance.

RocketQAv2: A Joint Training Method for Dense Passage Retrieval and Passage Re-ranking
Ruiyang Ren et al.

In various natural language processing tasks, passage retrieval and passage re-ranking are two key procedures in finding and ranking relevant information. Since both the two procedures contribute to the final performance, it is important to jointly optimize them in order to achieve mutual improvement. In this paper, we propose a novel joint training approach for dense passage retrieval and passage reranking. A major contribution is that we introduce the dynamic listwise distillation, where we design a unified listwise training approach for both the retriever and the re-ranker. During the dynamic distillation, the retriever and the re-ranker can be adaptively improved according to each other's relevance information. We also propose a hybrid data augmentation strategy to construct diverse training instances for listwise training approach. Extensive experiments show the effectiveness of our approach on both MSMARCO and Natural Questions datasets. Our code is available at https://github.com/PaddlePaddle/RocketQA.

Dealing with Typos for BERT-based Passage Retrieval and Ranking
Shengyao Zhuang and Guido Zuccon

Passage retrieval and ranking is a key task in open-domain question answering and information retrieval. Current effective approaches mostly rely on pre-trained deep language model-based retrievers and rankers. These methods have been shown to effectively model the semantic matching between queries and passages, also in presence of keyword mismatch, i.e. passages that are relevant to a query but do not contain important query keywords. In this paper we consider the Dense Retriever (DR), a passage retrieval method, and the BERT re-ranker, a popular passage re-ranking method. In this context, we formally investigate how these models respond and adapt to a specific type of keyword mismatch – that caused by keyword typos occurring in queries. Through empirical investigation, we find that typos can lead to a significant drop in retrieval and ranking effectiveness. We then propose a simple typos-aware training framework for DR and BERT re-ranker to address this issue. Our experimental results on the MS MARCO passage ranking dataset show that, with our proposed typos-aware training, DR and BERT re-ranker can become robust to typos in queries, resulting in significantly improved effectiveness compared to models trained without appropriately accounting for typos.

From Alignment to Assignment: Frustratingly Simple Unsupervised Entity Alignment
Xin Mao, Yuanbin Wu, and Man Lan

Cross-lingual entity alignment (EA) aims to find the equivalent entities between crosslingual KGs (Knowledge Graphs), which is a crucial step for integrating KGs. Recently, many GNN-based EA methods are proposed and show decent performance improvements on several public datasets. However, existing GNN-based EA methods inevitably inherit poor interpretability and low efficiency from neural networks. Motivated by the isomorphic assumption of GNN-based methods, we successfully transform the cross-lingual EA problem into an assignment problem. Based on this re-definition, we propose a frustratingly Simple but Effective Unsupervised entity alignment method (SEU) without neural networks. Extensive experiments have been conducted to show that our proposed unsupervised approach even beats advanced supervised methods across all public datasets while having high efficiency, interpretability, and stability.

Simple and Effective Unsupervised Redundancy Elimination to Compress Dense Vectors for Passage Retrieval
Xueguang Ma et al.

Recent work has shown that dense passage retrieval techniques achieve better ranking accuracy in open-domain question answering compared to sparse retrieval techniques such as BM25, but at the cost of large space and memory requirements. In this paper, we analyze the redundancy present in encoded dense vectors and show that the default dimension of 768 is unnecessarily large. To improve space efficiency, we propose a simple unsupervised compression pipeline that consists of principal component analysis (PCA), product quantization, and hybrid search. We further investigate other supervised baselines and find surprisingly that unsupervised PCA outperforms them in some settings. We perform extensive experiments on five question answering datasets and demonstrate that our best pipeline achieves good accuracy–space trade-offs, for example, $48\times$ compression with less than 3% drop in top-100 retrieval accuracy on average or $96\times$ compression with less than 4% drop. Code and data are available at `http://pyserini.io/`.

Relation Extraction with Word Graphs from N-grams
Han Qin, Yuanhe Tian, and Yan Song

Most recent studies for relation extraction (RE) leverage the dependency tree of the input sentence to incorporate syntax-driven contextual information to improve model performance, with little attention paid to the limitation where high-quality dependency parsers in most cases unavailable, especially for in-domain scenarios. To address this limitation, in this paper, we propose attentive graph convolutional networks (A-GCN) to improve neural RE methods with an unsupervised manner to build the context graph, without relying on the existence of a dependency parser. Specifically, we construct the graph from n-grams extracted from a lexicon built from pointwise mutual information (PMI) and apply attention over the graph. Therefore, different word pairs from the contexts within and across n-grams are weighted in the model and facilitate RE accordingly. Experimental results with further analyses on two English benchmark datasets for RE demonstrate the effectiveness of our approach, where state-of-the-art performance is observed on both datasets.

A Bayesian Framework for Information-Theoretic Probing
Tiago Pimentel and Ryan Cotterell

Pimentel et al. (2020) recently analysed probing from an information-theoretic perspective. They argue that probing should be seen as approximating a mutual information. This led to the rather unintuitive conclusion that representations encode exactly the same information about a target task as the original sentences. The mutual information, however, assumes the true probability distribution of a pair of random variables is known, leading to unintuitive results in settings where it is not. This paper proposes a new framework to measure what we term Bayesian mutual information, which analyses information from the perspective of Bayesian agents—allowing for more intuitive findings in scenarios with finite data. For instance, under Bayesian MI we have that data can add information, processing can help, and information can hurt, which makes it more intuitive for machine learning applications. Finally, we apply our framework to probing where we believe Bayesian mutual information naturally operationalises ease of extraction by explicitly limiting the available background knowledge to solve a task.

Masked Language Modeling and the Distributional Hypothesis: Order Word Matters Pre-training for Little
Koustuv Sinha et al.

A possible explanation for the impressive performance of masked language model (MLM) pre-training is that such models have learned to represent the syntactic structures prevalent in classical NLP pipelines. In this paper, we propose a different explanation: MLMs succeed on downstream tasks almost entirely due to their ability to model higher-order word co-occurrence statistics. To demonstrate this, we pre-train MLMs on sentences with randomly shuffled word order, and show that these models still achieve high accuracy after fine-tuning on many downstream tasks—including tasks specifically designed to be challenging for models that ignore word order. Our models perform surprisingly well according to some parametric syntactic probes, indicating possible deficiencies in how we test representations for syntactic information. Overall, our results show that purely distributional information largely explains the success of pre-training, and underscore the importance of curating challenging evaluation datasets that require deeper linguistic knowledge.

What's Hidden in a One-layer Randomly Weighted Transformer?
Sheng Shen et al.

We demonstrate that, hidden within one-layer randomly weighted neural networks, there exist subnetworks that can achieve impressive performance, without ever modifying the weight initializations, on machine translation tasks. To find subnetworks for one-layer randomly weighted neural networks, we apply different binary masks to the same weight matrix to generate different layers. Hidden within a one-layer randomly weighted Transformer, we find that subnetworks that can achieve 29.45/17.29 BLEU on IWSLT14/WMT14. Using a fixed pre-trained embedding layer, the previously found subnetworks are smaller than, but can match 98%/92% (34.14/25.24 BLEU) of the performance of, a trained Transformer$_small/base$ on IWSLT14/WMT14. Furthermore, we demonstrate the effectiveness of larger and deeper transformers in this setting, as well as the impact of different initialization methods.

Rethinking Denoised Auto-Encoding in Language Pre-Training
Fuli Luo et al.

Pre-trained self-supervised models such as BERT have achieved striking success in learning sequence representations, especially for natural language processing. These models typically corrupt the given sequences with certain types of noise, such as masking, shuffling, or substitution, and then try to recover the original input. However, such pre-training approaches are prone to learning representations that are covariant with the noise, leading to the discrepancy between the pre-training and fine-tuning stage. To remedy this, we present ContrAstive Pre-Training (CAPT) to learn noise invariant sequence representations. The proposed CAPT encourages the consistency between representations of the original sequence and its corrupted version via unsupervised instance-wise training signals. In this way, it not only alleviates the pretrain-finetune discrepancy induced by the noise of pre-training, but also aids the pre-trained model in better capturing global semantics of the input via more effective sentence-level supervision. Different from most prior work that focuses on a particular modality, comprehensive empirical evidence on 11 natural language understanding and cross-modal tasks illustrates that CAPT is applicable for both language and vision-language tasks, and obtains surprisingly consistent improvement, including 0.6% absolute gain on GLUE benchmarks and 0.8% absolute increment on NLVR2.

Lifelong Explainer for Lifelong Learners
Xuelin Situ et al.

Lifelong Learning (LL) black-box models are dynamic in that they keep learning from new tasks and constantly update their parameters. Owing to the need to utilize information from previously seen tasks, and capture commonalities in potentially diverse data, it is hard for automatic explanation methods to explain the outcomes of these models. In addition, existing explanation methods, e.g., LIME, which are computationally expensive when explaining a static black-box model, are even more inefficient in the LL setting. In this paper, we propose a novel Lifelong Explanation (LLE) approach that continuously trains a student explainer under the supervision of a teacher – an arbitrary explanation algorithm – on different tasks undertaken in LL. We also leverage the Experience Replay (ER) mechanism to prevent catastrophic forgetting in the student explainer. Our experiments comparing LLE to three baselines on text classification tasks show that LLE can enhance the stability of the explanations for all seen tasks and maintain the same level of faithfulness to the black-box model as the teacher, while being up to 10^2 times faster at test time. Our ablation study shows that the ER mechanism in our LLE approach enhances the learning capabilities of the student explainer. Our code is available at https://github.com/situsnow/LLE.

Linguistic Dependencies and Statistical Dependence
Jacob Louis Hoover et al.

Are pairs of words that tend to occur together also likely to stand in a linguistic dependency? This empirical question is motivated by a long history of literature in cognitive science, psycholinguistics, and NLP. In this work we contribute an extensive analysis of the relationship between linguistic dependencies and statistical dependence between words. Improving on previous work, we introduce the use of large pretrained language models to compute contextualized estimates of the pointwise mutual information between words (CPMI). For multiple models and languages, we extract dependency trees which maximize CPMI, and compare to gold standard linguistic dependencies. Overall, we find that CPMI dependencies achieve an unlabelled undirected attachment score of at most ≈ 0.5. While far above chance, and consistently above a non-contextualized PMI

baseline, this score is generally comparable to a simple baseline formed by connecting adjacent words. We analyze which kinds of linguistic dependencies are best captured in CPMI dependencies, and also find marked differences between the estimates of the large pretrained language models, illustrating how their different training schemes affect the type of dependencies they capture.

Modeling Human Sentence Processing with Left-Corner Recurrent Neural Network Grammars
Ryo Yoshida, Hiroshi Noji, and Yohei Oseki

In computational linguistics, it has been shown that hierarchical structures make language models (LMs) more human-like. However, the previous literature has been agnostic about a parsing strategy of the hierarchical models. In this paper, we investigated whether hierarchical structures make LMs more human-like, and if so, which parsing strategy is most cognitively plausible. In order to address this question, we evaluated three LMs against human reading times in Japanese with head-final left-branching structures: Long Short-Term Memory (LSTM) as a sequential model and Recurrent Neural Network Grammars (RNNGs) with top-down and left-corner parsing strategies as hierarchical models. Our computational modeling demonstrated that left-corner RNNGs outperformed top-down RNNGs and LSTM, suggesting that hierarchical and left-corner architectures are more cognitively plausible than top-down or sequential architectures. In addition, the relationships between the cognitive plausibility and (i) perplexity, (ii) parsing, and (iii) beam size will also be discussed.

A Simple and Effective Positional Encoding for Transformers
Pu-Chin Chen et al.

Transformer models are permutation equivariant. To supply the order and type information of the input tokens, position and segment embeddings are usually added to the input. Recent works proposed variations of positional encodings with relative position encodings achieving better performance. Our analysis shows that the gain actually comes from moving positional information to attention layer from the input. Motivated by this, we introduce Decoupled Positional Attention for Transformers (DIET), a simple yet effective mechanism to encode position and segment information into the Transformer models. The proposed method has faster training and inference time, while achieving competitive performance on GLUE, XTREME and WMT benchmarks. We further generalize our method to long-range transformers and show performance gain.

Explore Better Relative Position Embeddings from Encoding Perspective for Transformer Models
Anlin Qu, Jianwei Niu, and Shasha Mo

Relative position embedding (RPE) is a successful method to explicitly and efficaciously encode position information into Transformer models. In this paper, we investigate the potential problems in Shaw-RPE and XL-RPE, which are the most representative and prevalent RPEs, and propose two novel RPEs called Low-level Fine-grained High-level Coarse-grained (LFHC) RPE and Gaussian Cumulative Distribution Function (GCDF) RPE. LFHC-RPE is an improvement of Shaw-RPE, which enhances the perception ability at medium and long relative positions. GCDF-RPE utilizes the excellent properties of the Gaussian function to amend the prior encoding mechanism in XL-RPE. Experimental results on nine authoritative datasets demonstrate the effectiveness of our methods empirically. Furthermore, GCDF-RPE achieves the best overall performance among five different RPEs.

Adversarial Mixing Policy for Relaxing Locally Linear Constraints in Mixup
Guang Liu et al.

Mixup is a recent regularizer for current deep classification networks. Through training a neural network on convex combinations of pairs of examples and their labels, it imposes locally linear constraints on the model's input space. However, such strict linear constraints often lead to under-fitting which degrades the effects of regularization. Noticeably, this issue is getting more serious when the resource is extremely limited. To address these issues, we propose the Adversarial Mixing Policy (AMP), organized in a "min-max-rand" formulation, to relax the Locally Linear Constraints in Mixup. Specifically, AMP adds a small adversarial perturbation to the mixing coefficients rather than the examples. Thus, slight non-linearity is injected in-between the synthetic examples and synthetic labels. By training on these data, the deep networks are further regularized, and thus achieve a lower predictive error rate. Experiments on five text classification benchmarks and five backbone models have empirically shown that our methods reduce the error rate over Mixup variants in a significant margin (up to 31.3%), especially in low-resource conditions (up to 17.5%).

Is this the end of the gold standard? A straightforward reference-less grammatical error correction metric
Md Asadul Islam and Enrico Magnani

It is difficult to rank and evaluate the performance of grammatical error correction (GEC) systems, as a sentence can be rewritten in numerous correct ways. A number of GEC metrics have been used to evaluate proposed GEC systems; however, each system relies on either a comparison with one or more reference texts—in what is known as the gold standard for reference-based metrics—or a separate annotated dataset to fine-tune the reference-less metric. Reference-based systems have a low correlation with human judgement, cannot capture all the ways in which a sentence can be corrected, and require substantial work to develop a test dataset. We propose a reference-less GEC evaluation system that is strongly correlated with human judgement, solves the issues related to the use of a reference, and does not need another annotated dataset for fine-tuning. The

proposed system relies solely on commonly available tools. Additionally, currently available reference-less metrics do not work properly when part of a sentence is repeated as opposed to reference-based metrics. In our proposed system, we look to address issues inherent in reference-less metrics and reference-based metrics.

Augmenting BERT-style Models with Predictive Coding to Improve Discourse-level Representations

Vladimir Araujo et al.

Current language models are usually trained using a self-supervised scheme, where the main focus is learning representations at the word or sentence level. However, there has been limited progress in generating useful discourse-level representations. In this work, we propose to use ideas from predictive coding theory to augment BERT-style language models with a mechanism that allows them to learn suitable discourse-level representations. As a result, our proposed approach is able to predict future sentences using explicit top-down connections that operate at the intermediate layers of the network. By experimenting with benchmarks designed to evaluate discourse-related knowledge using pre-trained sentence representations, we demonstrate that our approach improves performance in 6 out of 11 tasks by excelling in discourse relationship detection.

Backdoor Attacks on Pre-trained Models by Layerwise Weight Poisoning

Linyang Li et al.

Pre-**T**rained **M**odels have been widely applied and recently proved vulnerable under backdoor attacks: the released pre-trained weights can be maliciously poisoned with certain triggers. When the triggers are activated, even the fine-tuned model will predict pre-defined labels, causing a security threat. These backdoors generated by the poisoning methods can be erased by changing hyper-parameters during fine-tuning or detected by finding the triggers. In this paper, we propose a stronger weight-poisoning attack method that introduces a layerwise weight poisoning strategy to plant deeper backdoors; we also introduce a combinatorial trigger that cannot be easily detected. The experiments on text classification tasks show that previous defense methods cannot resist our weight-poisoning method, which indicates that our method can be widely applied and may provide hints for future model robustness studies.

GAML-BERT: Improving BERT Early Exiting by Gradient Aligned Mutual Learning

Wei Zhu et al.

In this work, we propose a novel framework, Gradient Aligned Mutual Learning BERT (GAML-BERT), for improving the early exiting of BERT. GAML-BERT's contributions are two-fold. We conduct a set of pilot experiments, which shows that mutual knowledge distillation between a shallow exit and a deep exit leads to better performances for both. From this observation, we use mutual learning to improve BERT's early exiting performances, that is, we ask each exit of a multi-exit BERT to distill knowledge from each other. Second, we propose GA, a novel training method that aligns the gradients from knowledge distillation to cross-entropy losses. Extensive experiments are conducted on the GLUE benchmark, which shows that our GAML-BERT can significantly outperform the state-of-the-art (SOTA) BERT early exiting methods.

The Power of Scale for Parameter-Efficient Prompt Tuning

Brian Lester, Rami Al-Rfou, and Noah Constant

In this work, we explore "prompt tuning," a simple yet effective mechanism for learning "soft prompts" to condition frozen language models to perform specific downstream tasks. Unlike the discrete text prompts used by GPT-3, soft prompts are learned through backpropagation and can be tuned to incorporate signals from any number of labeled examples. Our end-to-end learned approach outperforms GPT-3's few-shot learning by a large margin. More remarkably, through ablations on model size using T5, we show that prompt tuning becomes more competitive with scale: as models exceed billions of parameters, our method "closes the gap" and matches the strong performance of model tuning (where all model weights are tuned). This finding is especially relevant because large models are costly to share and serve and the ability to reuse one frozen model for multiple downstream tasks can ease this burden. Our method can be seen as a simplification of the recently proposed "prefix tuning" of Li and Liang (2021) and we provide a comparison to this and other similar approaches. Finally, we show that conditioning a frozen model with soft prompts confers benefits in robustness to domain transfer and enables efficient "prompt ensembling." We release code and model checkpoints to reproduce our experiments.

Scalable Font Reconstruction with Dual Latent Manifolds

Nikita Srivatsan et al.

We propose a deep generative model that performs typography analysis and font reconstruction by learning disentangled manifolds of both font style and character shape. Our approach enables us to massively scale up the number of character types we can effectively model compared to previous methods. Specifically, we infer separate latent variables representing character and font via a pair of inference networks which take as input sets of glyphs that either all share a character type, or belong to the same font. This design allows our model to generalize to characters that were not observed during training time, an important task in light of the relative sparsity of most fonts. We also put forward a new loss, adapted from prior work that measures likelihood using an adaptive distribution in a projected space, resulting in more natural images without requiring a discriminator. We evaluate on the task of font reconstruction over various datasets representing character types of many languages, and compare favorably to modern style transfer systems according to both automatic and manually-evaluated metrics.

Neuro-Symbolic Approaches for Text-Based Policy Learning
Subhajit Chaudhury et al.

Text-Based Games (TBGs) have emerged as important testbeds for reinforcement learning (RL) in the natural language domain. Previous methods using LSTM-based action policies are uninterpretable and often overfit the training games showing poor performance to unseen test games. We present SymboLic Action policy for Textual Environments (SLATE), that learns interpretable action policy rules from symbolic abstractions of textual observations for improved generalization. We outline a method for end-to-end differentiable symbolic rule learning and show that such symbolic policies outperform previous state-of-the-art methods in text-based RL for the coin collector environment from 5-10x fewer training games. Additionally, our method provides human-understandable policy rules that can be readily verified for their logical consistency and can be easily debugged.

Layer-wise Model Pruning based on Mutual Information
Chun Fan et al.

Inspired by mutual information (MI) based feature selection in SVMs and logistic regression, in this paper, we propose MI-based layer-wise pruning: for each layer of a multi-layer neural network, neurons with higher values of MI with respect to preserved neurons in the upper layer are preserved. Starting from the top softmax layer, layer-wise pruning proceeds in a top-down fashion until reaching the bottom word embedding layer. The proposed pruning strategy offers merits over weight-based pruning techniques: (1) it avoids irregular memory access since representations and matrices can be squeezed into their smaller but dense counterparts, leading to greater speedup; (2) in a manner of top-down pruning, the proposed method operates from a more global perspective based on training signals in the top layer, and prunes each layer by propagating the effect of global signals through layers, leading to better performances at the same sparsity level. Extensive experiments show that at the same sparsity level, the proposed strategy offers both greater speedup and higher performances than weight-based pruning methods (e.g., magnitude pruning, movement pruning).

Hierarchical Heterogeneous Graph Representation Learning for Short Text Classification
Yaqing Wang et al.

Short text classification is a fundamental task in natural language processing. It is hard due to the lack of context information and labeled data in practice. In this paper, we propose a new method called SHINE, which is based on graph neural network (GNN), for short text classification. First, we model the short text dataset as a hierarchical heterogeneous graph consisting of word-level component graphs which introduce more semantic and syntactic information. Then, we dynamically learn a short document graph that facilitates effective label propagation among similar short texts. Thus, comparing with existing GNN-based methods, SHINE can better exploit interactions between nodes of the same types and capture similarities between short texts. Extensive experiments on various benchmark short text datasets show that SHINE consistently outperforms state-of-the-art methods, especially with fewer labels.

kFolden: k-Fold Ensemble for Out-Of-Distribution Detection
Xiaoya Li et al.

Out-of-Distribution (OOD) detection is an important problem in natural language processing (NLP). In this work, we propose a simple yet effective framework kFolden, which mimics the behaviors of OOD detection during training without the use of any external data. For a task with k training labels, kFolden induces k sub-models, each of which is trained on a subset with $k-1$ categories with the left category masked unknown to the sub-model. Exposing an unknown label to the sub-model during training, the model is encouraged to learn to equally attribute the probability to the seen $k-1$ labels for the unknown label, enabling this framework to simultaneously resolve in- and out-distribution examples in a natural way via OOD simulations. Taking text classification as an archetype, we develop benchmarks for OOD detection using existing text classification datasets. By conducting comprehensive comparisons and analyses on the developed benchmarks, we demonstrate the superiority of kFolden against current methods in terms of improving OOD detection performances while maintaining improved in-domain classification accuracy.

Frustratingly Simple Pretraining Alternatives to Masked Language Modeling
Atsuki Yamaguchi et al.

Masked language modeling (MLM), a self-supervised pretraining objective, is widely used in natural language processing for learning text representations. MLM trains a model to predict a random sample of input tokens that have been replaced by a [MASK] placeholder in a multi-class setting over the entire vocabulary. When pretraining, it is common to use alongside MLM other auxiliary objectives on the token or sequence level to improve downstream performance (e.g. next sentence prediction). However, no previous work so far has attempted in examining whether other simpler linguistically intuitive or not objectives can be used standalone as main pretraining objectives. In this paper, we explore five simple pretraining objectives based on token-level classification tasks as replacements of MLM. Empirical results on GLUE and SQUAD show that our proposed methods achieve comparable or better performance to MLM using a BERT-BASE architecture. We further validate our methods using smaller models, showing that pretraining a model with 41% of the BERT-BASE's parameters, BERT-MEDIUM results in only a 1% drop in GLUE scores with our best objective.

HRKD: Hierarchical Relational Knowledge Distillation for Cross-domain Language Model Compression
Chenhe Dong et al.

On many natural language processing tasks, large pre-trained language models (PLMs) have shown overwhelming performances compared with traditional neural network methods. Nevertheless, their huge model size and low inference speed have hindered the deployment on resource-limited devices in practice. In this paper, we target to compress PLMs with knowledge distillation, and propose a hierarchical relational knowledge distillation (HRKD) method to capture both hierarchical and domain relational information. Specifically, to enhance the model capability and transferability, we leverage the idea of meta-learning and set up domain-relational graphs to capture the relational information across different domains. And to dynamically select the most representative prototypes for each domain, we propose a hierarchical compare-aggregate mechanism to capture hierarchical relationships. Extensive experiments on public multi-domain datasets demonstrate the superior performance of our HRKD method as well as its strong few-shot learning ability. For reproducibility, we release the code at `https://github.com/cheneydon/hrkd`.

Searching for an Effective Defender: Benchmarking Defense against Adversarial Word Substitution
Zongyi Li et al.

Recent studies have shown that deep neural network-based models are vulnerable to intentionally crafted adversarial examples, and various methods have been proposed to defend against adversarial word-substitution attacks for neural NLP models. However, there is a lack of systematic study on comparing different defense approaches under the same attacking setting. In this paper, we seek to fill the gap of systematic studies through comprehensive researches on understanding the behavior of neural text classifiers trained by various defense methods under representative adversarial attacks. In addition, we propose an effective method to further improve the robustness of neural text classifiers against such attacks, and achieved the highest accuracy on both clean and adversarial examples on AGNEWS and IMDB datasets by a significant margin. We hope this study could provide useful clues for future research on text adversarial defense. Codes are available at https://github.com/RockyLzy/TextDefender.

Re-embedding Difficult Samples via Mutual Information Constrained Semantically Oversampling for Imbalanced Text Classification
Jiachen Tian et al.

Difficult samples of the minority class in imbalanced text classification are usually hard to be classified as they are embedded into an overlapping semantic region with the majority class. In this paper, we propose a Mutual Information constrained Semantically Oversampling framework (MISO) that can generate anchor instances to help the backbone network determine the re-embedding position of a non-overlapping representation for each difficult sample. MISO consists of (1) a semantic fusion module that learns entangled semantics among difficult and majority samples with an adaptive multi-head attention mechanism, (2) a mutual information loss that forces our model to learn new representations of entangled semantics in the non-overlapping region of the minority class, and (3) a coupled adversarial encoder-decoder that fine-tunes disentangled semantic representations to remain their correlations with the minority class, and then using these disentangled semantic representations to generate anchor instances for each difficult sample. Experiments on a variety of imbalanced text classification tasks demonstrate that anchor instances help classifiers achieve significant improvements over strong baselines.

Beyond Text: Incorporating Metadata and Label Structure for Multi-Label Document Classification using Heterogeneous Graphs
Chenchen Ye et al.

Multi-label document classification, associating one document instance with a set of relevant labels, is attracting more and more research attention. Existing methods explore the incorporation of information beyond text, such as document metadata or label structure. These approaches however either simply utilize the semantic information of metadata or employ the predefined parent-child label hierarchy, ignoring the heterogeneous graphical structures of metadata and labels, which we believe are crucial for accurate multi-label document classification. Therefore, in this paper, we propose a novel neural network based approach for multi-label document classification, in which two heterogeneous graphs are constructed and learned using heterogeneous graph transformers. One is metadata heterogeneous graph, which models various types of metadata and their topological relations. The other is label heterogeneous graph, which is constructed based on both the labels' hierarchy and their statistical dependencies. Experimental results on two benchmark datasets show the proposed approach outperforms several state-of-the-art baselines.

Natural Language Processing Meets Quantum Physics: A Survey and Categorization
Sixuan Wu et al.

Recent research has investigated quantum NLP, designing algorithms that process natural language in quantum computers, and also quantum-inspired algorithms that improve NLP performance on classical computers. In this survey, we review representative methods at the intersection of NLP and quantum physics in the past ten years, categorizing them according to the use of quantum theory, the linguistic targets that are modeled, and the downstream application. The literature review ends with a discussion on the key factors to the success that has been achieved by existing work, as well as challenges ahead, with the goal of better understanding the promises and further directions.

MetaTS: Meta Teacher-Student Network for Multilingual Sequence Labeling with Minimal Supervision
Zheng Li et al.

Sequence labeling aims to predict a fine-grained sequence of labels for the text. However, such formulation hinders the effectiveness of supervised methods due to the lack of token-level annotated data. This is exacerbated when we meet a diverse range of languages. In this work, we explore multilingual sequence labeling with minimal supervision using a single unified model for multiple languages. Specifically, we propose a Meta Teacher-Student (MetaTS) Network, a novel meta learning method to alleviate data scarcity by leveraging large multilingual unlabeled data. Prior teacher-student frameworks of self-training rely on rigid teaching strategies, which may hardly produce high-quality pseudo-labels for consecutive and interdependent tokens. On the contrary, MetaTS allows the teacher to dynamically adapt its pseudo-annotation strategies by the student's feedback on the generated pseudo-labeled data of each language and thus mitigate error propagation from noisy pseudo-labels. Extensive experiments on both public and real-world multilingual sequence labeling datasets empirically demonstrate the effectiveness of MetaTS.

Neural Machine Translation with Heterogeneous Topic Knowledge Embeddings
Weixuan Wang et al.

Neural Machine Translation (NMT) has shown a strong ability to utilize local context to disambiguate the meaning of words. However, it remains a challenge for NMT to leverage broader context information like topics. In this paper, we propose heterogeneous ways of embedding topic information at the sentence level into an NMT model to improve translation performance. Specifically, the topic information can be incorporated as pre-encoder topic embedding, post-encoder topic embedding, and decoder topic embedding to increase the likelihood of selecting target words from the same topic of the source sentence. Experimental results show that NMT models with the proposed topic knowledge embedding outperform the baselines on the English -> German and English -> French translation tasks.

Allocating Large Vocabulary Capacity for Cross-Lingual Language Model Pre-Training
Bo Zheng et al.

Compared to monolingual models, cross-lingual models usually require a more expressive vocabulary to represent all languages adequately. We find that many languages are under-represented in recent cross-lingual language models due to the limited vocabulary capacity. To this end, we propose an algorithm VoCap to determine the desired vocabulary capacity of each language. However, increasing the vocabulary size significantly slows down the pre-training speed. In order to address the issues, we propose k-NN-based target sampling to accelerate the expensive softmax. Our experiments show that the multilingual vocabulary learned with VoCap benefits cross-lingual language model pre-training. Moreover, k-NN-based target sampling mitigates the side-effects of increasing the vocabulary size while achieving comparable performance and faster pre-training speed. The code and the pretrained multilingual vocabularies are available at https://github.com/bozheng-hit/VoCapXLM.

Recurrent Attention for Neural Machine Translation
Jiali Zeng et al.

Recent research questions the importance of the dot-product self-attention in Transformer models and shows that most attention heads learn simple positional patterns. In this paper, we push further in this research line and propose a novel substitute mechanism for self-attention: Recurrent AtteNtion (RAN) . RAN directly learns attention weights without any token-to-token interaction and further improves their capacity by layer-to-layer interaction. Across an extensive set of experiments on 10 machine translation tasks, we find that RAN models are competitive and outperform their Transformer counterpart in certain scenarios, with fewer parameters and inference time. Particularly, when apply RAN to the decoder of Transformer, there brings consistent improvements by about +0.5 BLEU on 6 translation tasks and +1.0 BLEU on Turkish-English translation task. In addition, we conduct extensive analysis on the attention weights of RAN to confirm their reasonableness. Our RAN is a promising alternative to build more effective and efficient NMT models.

Learning from Multiple Noisy Augmented Data Sets for Better Cross-Lingual Spoken Language Understanding
Yingmei Guo et al.

Lack of training data presents a grand challenge to scaling out spoken language understanding (SLU) to low-resource languages. Although various data augmentation approaches have been proposed to synthesize training data in low-resource target languages, the augmented data sets are often noisy, and thus impede the performance of SLU models. In this paper we focus on mitigating noise in augmented data. We develop a denoising training approach. Multiple models are trained with data produced by various augmented methods. Those models provide supervision signals to each other. The experimental results show that our method outperforms the existing state of the art by 3.05 and 4.24 percentage points on two benchmark datasets, respectively. The code will be made open sourced on github.

Enlivening Redundant Heads in Multi-head Self-attention for Machine Translation
Tianfu Zhang et al.

Multi-head self-attention recently attracts enormous interest owing to its specialized functions, significant parallelizable computation, and flexible extensibility. However, very recent empirical studies show that some

self-attention heads make little contribution and can be pruned as redundant heads. This work takes a novel perspective of identifying and then vitalizing redundant heads. We propose a redundant head enlivening (RHE) method to precisely identify redundant heads, and then vitalize their potential by learning syntactic relations and prior knowledge in the text without sacrificing the roles of important heads. Two novel syntax-enhanced attention (SEA) mechanisms: a dependency mask bias and a relative local-phrasal position bias, are introduced to revise self-attention distributions for syntactic enhancement in machine translation. The importance of individual heads is dynamically evaluated during the redundant heads identification, on which we apply SEA to vitalize redundant heads while maintaining the strength of important heads. Experimental results on widely adopted WMT14 and WMT16 English to German and English to Czech language machine translation validate the RHE effectiveness.

Unsupervised Neural Machine Translation with Universal Grammar
Zuchao Li et al.

Machine translation usually relies on parallel corpora to provide parallel signals for training. The advent of unsupervised machine translation has brought machine translation away from this reliance, though performance still lags behind traditional supervised machine translation. In unsupervised machine translation, the model seeks symmetric language similarities as a source of weak parallel signal to achieve translation. Chomsky's Universal Grammar theory postulates that grammar is an innate form of knowledge to humans and is governed by universal principles and constraints. Therefore, in this paper, we seek to leverage such shared grammar clues to provide more explicit language parallel signals to enhance the training of unsupervised machine translation models. Through experiments on multiple typical language pairs, we demonstrate the effectiveness of our proposed approaches.

Encouraging Lexical Translation Consistency for Document-Level Neural Machine Translation
Xinglin Lyu et al.

Recently a number of approaches have been proposed to improve translation performance for document-level neural machine translation (NMT). However, few are focusing on the subject of lexical translation consistency. In this paper we apply "one translation per discourse" in NMT, and aim to encourage lexical translation consistency for document-level NMT. This is done by first obtaining a word link for each source word in a document, which tells the positions where the source word appears. Then we encourage the translation of those words within a link to be consistent in two ways. On the one hand, when encoding sentences within a document we properly share context information of those words. On the other hand, we propose an auxiliary loss function to better constrain that their translation should be consistent. Experimental results on ChineseEnglish and English→French translation tasks show that our approach not only achieves state-of-the-art performance in BLEU scores, but also greatly improves lexical consistency in translation.

Improving Neural Machine Translation by Bidirectional Training
Liang Ding, Di Wu, and Dacheng Tao

We present a simple and effective pretraining strategy – bidirectional training (BiT) for neural machine translation. Specifically, we bidirectionally update the model parameters at the early stage and then tune the model normally. To achieve bidirectional updating, we simply reconstruct the training samples from "src→tgt" to "src+tgt→tgt+src" without any complicated model modifications. Notably, our approach does not increase any parameters or training steps, requiring the parallel data merely. Experimental results show that BiT pushes the SOTA neural machine translation performance across 15 translation tasks on 8 language pairs (data sizes range from 160K to 38M) significantly higher. Encouragingly, our proposed model can complement existing data manipulation strategies, i.e. back translation, data distillation, and data diversification. Extensive analyses show that our approach functions as a novel bilingual code-switcher, obtaining better bilingual alignment.

Scheduled Sampling Based on Decoding Steps for Neural Machine Translation
Yijin Liu et al.

Scheduled sampling is widely used to mitigate the exposure bias problem for neural machine translation. Its core motivation is to simulate the inference scene during training by replacing ground-truth tokens with predicted tokens, thus bridging the gap between training and inference. However, vanilla scheduled sampling is merely based on training steps and equally treats all decoding steps. Namely, it simulates an inference scene with uniform error rates, which disobeys the real inference scene, where larger decoding steps usually have higher error rates due to error accumulations. To alleviate the above discrepancy, we propose scheduled sampling methods based on decoding steps, increasing the selection chance of predicted tokens with the growth of decoding steps. Consequently, we can more realistically simulate the inference scene during training, thus better bridging the gap between training and inference. Moreover, we investigate scheduled sampling based on both training steps and decoding steps for further improvements. Experimentally, our approaches significantly outperform the Transformer baseline and vanilla scheduled sampling on three large-scale WMT tasks. Additionally, our approaches also generalize well to the text summarization task on two popular benchmarks.

Learning to Rewrite for Non-Autoregressive Neural Machine Translation
Xinwei Geng, Xiaocheng Feng, and Bing Qin

Non-autoregressive neural machine translation, which decomposes the dependence on previous target tokens from the inputs of the decoder, has achieved impressive inference speedup but at the cost of inferior accuracy.

Previous works employ iterative decoding to improve the translation by applying multiple refinement iterations. However, a serious drawback is that these approaches expose the serious weakness in recognizing the erroneous translation pieces. In this paper, we propose an architecture named RewriteNAT to explicitly learn to rewrite the erroneous translation pieces. Specifically, RewriteNAT utilizes a locator module to locate the erroneous ones, which are then revised into the correct ones by a revisor module. Towards keeping the consistency of data distribution with iterative decoding, an iterative training strategy is employed to further improve the capacity of rewriting. Extensive experiments conducted on several widely-used benchmarks show that RewriteNAT can achieve better performance while significantly reducing decoding time, compared with previous iterative decoding strategies. In particular, RewriteNAT can obtain competitive results with autoregressive translation on WMT14 En-De, En-Fr and WMT16 Ro-En translation benchmarks.

SHAPE: Shifted Absolute Position Embedding for Transformers
Shun Kiyono et al.

Position representation is crucial for building position-aware representations in Transformers. Existing position representations suffer from a lack of generalization to test data with unseen lengths or high computational cost. We investigate shifted absolute position embedding (SHAPE) to address both issues. The basic idea of SHAPE is to achieve shift invariance, which is a key property of recent successful position representations, by randomly shifting absolute positions during training. We demonstrate that SHAPE is empirically comparable to its counterpart while being simpler and faster.

Self-Supervised Quality Estimation for Machine Translation
Yuanhang Zheng et al.

Quality estimation (QE) of machine translation (MT) aims to evaluate the quality of machine-translated sentences without references and is important in practical applications of MT. Training QE models require massive parallel data with hand-crafted quality annotations, which are time-consuming and labor-intensive to obtain. To address the issue of the absence of annotated training data, previous studies attempt to develop unsupervised QE methods. However, very few of them can be applied to both sentence- and word-level QE tasks, and they may suffer from noises in the synthetic data. To reduce the negative impact of noises, we propose a self-supervised method for both sentence- and word-level QE, which performs quality estimation by recovering the masked target words. Experimental results show that our method outperforms previous unsupervised methods on several QE tasks in different language pairs and domains.

Generalised Unsupervised Domain Adaptation of Neural Machine Translation with Cross-Lingual Data Selection
Thuy-Trang Vu et al.

This paper considers the unsupervised domain adaptation problem for neural machine translation (NMT), where we assume the access to only monolingual text in either the source or target language in the new domain. We propose a cross-lingual data selection method to extract in-domain sentences in the missing language side from a large generic monolingual corpus. Our proposed method trains an adaptive layer on top of multilingual BERT by contrastive learning to align the representation between the source and target language. This then enables the transferability of the domain classifier between the languages in a zero-shot manner. Once the in-domain data is detected by the classifier, the NMT model is then adapted to the new domain by jointly learning translation and domain discrimination tasks. We evaluate our cross-lingual data selection method on NMT across five diverse domains in three language pairs, as well as a real-world scenario of translation for COVID-19. The results show that our proposed method outperforms other selection baselines up to +1.5 BLEU score.

STANKER: Stacking Network based on Level-grained Attention-masked BERT for Rumor Detection on Social Media
Dongning Rao et al.

Rumor detection on social media puts pre-trained language models (LMs), such as BERT, and auxiliary features, such as comments, into use. However, on the one hand, rumor detection datasets in Chinese companies with comments are rare; on the other hand, intensive interaction of attention on Transformer-based models like BERT may hinder performance improvement. To alleviate these problems, we build a new Chinese microblog dataset named Weibo20 by collecting posts and associated comments from Sina Weibo and propose a new ensemble named STANKER (Stacking neTwork bAsed-on atteNtion-masKed BERT). STANKER adopts two level-grained attention-masked BERT (LGAM-BERT) models as base encoders. Unlike the original BERT, our new LGAM-BERT model takes comments as important auxiliary features and masks co-attention between posts and comments on lower-layers. Experiments on Weibo20 and three existing social media datasets showed that STANKER outperformed all compared models, especially beating the old state-of-the-art on Weibo dataset.

ActiveEA: Active Learning for Neural Entity Alignment
Bing Liu et al.

Entity Alignment (EA) aims to match equivalent entities across different Knowledge Graphs (KGs) and is an essential step of KG fusion. Current mainstream methods – neural EA models – rely on training with seed alignment, i.e., a set of pre-aligned entity pairs which are very costly to annotate. In this paper, we devise a novel Active Learning (AL) framework for neural EA, aiming to create highly informative seed

alignment to obtain more effective EA models with less annotation cost. Our framework tackles two main challenges encountered when applying AL to EA: (1) How to exploit dependencies between entities within the AL strategy. Most AL strategies assume that the data instances to sample are independent and identically distributed. However, entities in KGs are related. To address this challenge, we propose a structure-aware uncertainty sampling strategy that can measure the uncertainty of each entity as well as its impact on its neighbour entities in the KG. (2) How to recognise entities that appear in one KG but not in the other KG (i.e., bachelors). Identifying bachelors would likely save annotation budget. To address this challenge, we devise a bachelor recognizer paying attention to alleviate the effect of sampling bias. Empirical results show that our proposed AL strategy can significantly improve sampling quality with good generality across different datasets, EA models and amount of bachelors.

Cost-effective End-to-end Information Extraction for Semi-structured Document Images

Wonseok Hwang et al.

A real-world information extraction (IE) system for semi-structured document images often involves a long pipeline of multiple modules, whose complexity dramatically increases its development and maintenance cost. One can instead consider an end-to-end model that directly maps the input to the target output and simplify the entire process. However, such generation approach is known to lead to unstable performance if not designed carefully. Here we present our recent effort on transitioning from our existing pipeline-based IE system to an end-to-end system focusing on practical challenges that are associated with replacing and deploying the system in real, large-scale production. By carefully formulating document IE as a sequence generation task, we show that a single end-to-end IE system can be built and still achieve competent performance.

Improving Math Word Problems with Pre-trained Knowledge and Hierarchical Reasoning

Weijiang Yu et al.

The recent algorithms for math word problems (MWP) neglect to use outside knowledge not present in the problems. Most of them only capture the word-level relationship and ignore to build hierarchical reasoning like the human being for mining the contextual structure between words and sentences. In this paper, we propose a **R**easoning with **P**re-trained **K**nowledge and **H**ierarchical **S**tructure (**RPKHS**) network, which contains a pre-trained knowledge encoder and a hierarchical reasoning encoder. Firstly, our pre-trained knowledge encoder aims at reasoning the MWP by using outside knowledge from the pre-trained transformer-based models. Secondly, the hierarchical reasoning encoder is presented for seamlessly integrating the word-level and sentence-level reasoning to bridge the entity and context domain on MWP. Extensive experiments show that our RPKHS significantly outperforms state-of-the-art approaches on two large-scale commonly-used datasets, and boosts performance from 77.4% to 83.9% on Math23K, from 75.5 to 82.2% on Math23K with 5-fold cross-validation and from 83.7% to 89.8% on MAWPS. More extensive ablations are shown to demonstrate the effectiveness and interpretability of our proposed method.

GraphMR: Graph Neural Network for Mathematical Reasoning

Weijie Feng et al.

Mathematical reasoning aims to infer satisfiable solutions based on the given mathematics questions. Previous natural language processing researches have proven the effectiveness of sequence-to-sequence (Seq2Seq) or related variants on mathematics solving. However, few works have been able to explore structural or syntactic information hidden in expressions (e.g., precedence and associativity). This dissertation set out to investigate the usefulness of such untapped information for neural architectures. Firstly, mathematical questions are represented in the format of graphs within syntax analysis. The structured nature of graphs allows them to represent relations of variables or operators while preserving the semantics of the expressions. Having transformed to the new representations, we proposed a graph-to-sequence neural network GraphMR, which can effectively learn the hierarchical information of graphs inputs to solve mathematics and speculate answers. A complete experimental scenario with four classes of mathematical tasks and three Seq2Seq baselines is built to conduct a comprehensive analysis, and results show that GraphMR outperforms others in hidden information learning and mathematics resolving.

What Changes Can Large-scale Language Models Bring? Intensive Study on HyperCLOVA: Billions-scale Korean Generative Pretrained Transformers

Boseop Kim et al.

GPT-3 shows remarkable in-context learning ability of large-scale language models (LMs) trained on hundreds of billion scale data. Here we address some remaining issues less reported by the GPT-3 paper, such as a non-English LM, the performances of different sized models, and the effect of recently introduced prompt optimization on in-context learning. To achieve this, we introduce HyperCLOVA, a Korean variant of 82B GPT-3 trained on a Korean-centric corpus of 560B tokens. Enhanced by our Korean-specific tokenization, HyperCLOVA with our training configuration shows state-of-the-art in-context zero-shot and few-shot learning performances on various downstream tasks in Korean. Also, we show the performance benefits of prompt-based learning and demonstrate how it can be integrated into the prompt engineering pipeline. Then we discuss the possibility of materializing the No Code AI paradigm by providing AI prototyping capabilities to non-experts of ML by introducing HyperCLOVA studio, an interactive prompt engineering interface. Lastly, we demonstrate the potential of our methods with three successful in-house applications.

APIRecX: Cross-Library API Recommendation via Pre-Trained Language Model

Yuning Kang et al.

For programmers, learning the usage of APIs (Application Programming Interfaces) of a software library is important yet difficult. API recommendation tools can help developers use APIs by recommending which APIs to be used next given the APIs that have been written. Traditionally, language models such as N-gram are applied to API recommendation. However, because the software libraries keep changing and new libraries keep emerging, new APIs are common. These new APIs can be seen as OOV (out of vocabulary) words and cannot be handled well by existing API recommendation approaches due to the lack of training data. In this paper, we propose APIRecX, the first cross-library API recommendation approach, which uses BPE to split each API call in each API sequence and pre-trains a GPT based language model. It then recommends APIs by fine-tuning the pre-trained model. APIRecX can migrate the knowledge of existing libraries to a new library, and can recommend APIs that are previously regarded as OOV. We evaluate APIRecX on six libraries and the results confirm its effectiveness by comparing with two typical API recommendation approaches.

GMH: A General Multi-hop Reasoning Model for KG Completion

Yao Zhang et al.

Knowledge graphs are essential for numerous downstream natural language processing applications, but are typically incomplete with many facts missing. This results in research efforts on multi-hop reasoning task, which can be formulated as a search process and current models typically perform short distance reasoning. However, the long-distance reasoning is also vital with the ability to connect the superficially unrelated entities. To the best of our knowledge, there lacks a general framework that approaches multi-hop reasoning in mixed long-short distance reasoning scenarios. We argue that there are two key issues for a general multi-hop reasoning model: i) where to go, and ii) when to stop. Therefore, we propose a general model which resolves the issues with three modules: 1) the local-global knowledge module to estimate the possible paths, 2) the differentiated action dropout module to explore a diverse set of paths, and 3) the adaptive stopping search module to avoid over searching. The comprehensive results on three datasets demonstrate the superiority of our model with significant improvements against baselines in both short and long distance reasoning scenarios.

BPM_MT: Enhanced Backchannel Prediction Model using Multi-Task Learning

Jin Yea Jang et al.

Backchannel (BC), a short reaction signal of a listener to a speaker's utterances, helps to improve the quality of the conversation. Several studies have been conducted to predict BC in conversation; however, the utilization of advanced natural language processing techniques using lexical information presented in the utterances of a speaker has been less considered. To address this limitation, we present a BC prediction model called BPM_MT (Backchannel prediction model with multitask learning), which utilizes KoBERT, a pre-trained language model. The BPM_MT simultaneously carries out two tasks at learning: 1) BC category prediction using acoustic and lexical features, and 2) sentiment score prediction based on sentiment cues. BPM_MT exhibited 14.24% performance improvement compared to the existing baseline in the four BC categories: continuer, understanding, empathic response, and No BC. In particular, for empathic response category, a performance improvement of 17.14% was achieved.

Graphine: A Dataset for Graph-aware Terminology Definition Generation

Zequn Liu et al.

Precisely defining the terminology is the first step in scientific communication. Developing neural text generation models for definition generation can circumvent the labor-intensity curation, further accelerating scientific discovery. Unfortunately, the lack of large-scale terminology definition dataset hinders the process toward definition generation. In this paper, we present a large-scale terminology definition dataset Graphine covering 2,010,648 terminology definition pairs, spanning 227 biomedical subdisciplines. Terminologies in each subdiscipline further form a directed acyclic graph, opening up new avenues for developing graph-aware text generation models. We then proposed a novel graph-aware definition generation model Graphex that integrates transformer with graph neural network. Our model outperforms existing text generation models by exploiting the graph structure of terminologies. We further demonstrated how Graphine can be used to evaluate pre-trained language models, compare graph representation learning methods and predict sentence granularity. We envision Graphine to be a unique resource for definition generation and many other NLP tasks in biomedicine.

Leveraging Order-Free Tag Relations for Context-Aware Recommendation

Junmo Kang et al.

Tag recommendation relies on either a ranking function for top-k tags or an autoregressive generation method. However, the previous methods neglect one of two seemingly conflicting yet desirable characteristics of a tag set: orderlessness and inter-dependency. While the ranking approach fails to address the inter-dependency among tags when they are ranked, the autoregressive approach fails to take orderlessness into account because it is designed to utilize sequential relations among tokens. We propose a sequence-oblivious generation method for tag recommendation, in which the next tag to be generated is independent of the order of the generated tags and the order of the ground truth tags occurring in training data. Empirical results on two different domains, Instagram and Stack Overflow, show that our method is significantly superior to the previous approaches.

End-to-End Conversational Search for Online Shopping with Utterance Transfer

Liqiang Xiao et al.

Successful conversational search systems can present natural, adaptive and interactive shopping experience for online shopping customers. However, building such systems from scratch faces real word challenges

from both imperfect product schema/knowledge and lack of training dialog data. In this work we first propose ConvSearch, an end-to-end conversational search system that deeply combines the dialog system with search. It leverages the text profile to retrieve products, which is more robust against imperfect product schema/knowledge compared with using product attributes alone. We then address the lack of data challenges by proposing an utterance transfer approach that generates dialogue utterances by using existing dialog from other domains, and leveraging the search behavior data from e-commerce retailer. With utterance transfer, we introduce a new conversational search dataset for online shopping. Experiments show that our utterance transfer method can significantly improve the availability of training dialogue data without crowd-sourcing, and the conversational search system significantly outperformed the best tested baseline.

Self-Supervised Curriculum Learning for Spelling Error Correction
Zifa Gan, Hongfei Xu, and Hongying Zan

Spelling Error Correction (SEC) that requires high-level language understanding is a challenging but useful task. Current SEC approaches normally leverage a pre-training then fine-tuning procedure that treats data equally. By contrast, Curriculum Learning (CL) utilizes training data differently during training and has shown its effectiveness in improving both performance and training efficiency in many other NLP tasks. In NMT, a model's performance has been shown sensitive to the difficulty of training examples, and CL has been shown effective to address this. In SEC, the data from different language learners are naturally distributed at different difficulty levels (some errors made by beginners are obvious to correct while some made by fluent speakers are hard), and we expect that designing a curriculum correspondingly for model learning may also help its training and bring about better performance. In this paper, we study how to further improve the performance of the state-of-the-art SEC method with CL, and propose a Self-Supervised Curriculum Learning (SSCL) approach. Specifically, we directly use the cross-entropy loss as criteria for: 1) scoring the difficulty of training data, and 2) evaluating the competence of the model. In our approach, CL improves the model training, which in return improves the CL measurement. In our experiments on the SIGHAN 2015 Chinese spelling check task, we show that SSCL is superior to previous norm-based and uncertainty-aware approaches, and establish a new state of the art (74.38% F1).

Fix-Filter-Fix: Intuitively Connect Any Models for Effective Bug Fixing
Haiwen Hong et al.

Locating and fixing bugs is a time-consuming task. Most neural machine translation (NMT) based approaches for automatically bug fixing lack generality and do not make full use of the rich information in the source code. In NMT-based bug fixing, we find some predicted code identical to the input buggy code (called unchanged fix) in NMT-based approaches due to high similarity between buggy and fixed code (e.g., the difference may only appear in one particular line). Obviously, unchanged fix is not the correct fix because it is the same as the buggy code that needs to be fixed. Based on these, we propose an intuitive yet effective general framework (called Fix-Filter-Fix or F^3) for bug fixing. F^3 connects models with our filter mechanism to filter out the last model's unchanged fix to the next. We propose an F^3 theory that can quantitatively and accurately calculate the F^3 lifting effect. To evaluate, we implement the Seq2Seq Transformer (ST) and the AST2Seq Transformer (AT) to form some basic F^3 instances, called $F^3\_ST + AT$ and $F^3\_AT + ST$. Comparing them with single model approaches and many model connection baselines across four datasets validates the effectiveness and generality of F^3 and corroborates our findings and methodology.

Neuro-Symbolic Reinforcement Learning with First-Order Logic
Daiki Kimura et al.

Deep reinforcement learning (RL) methods often require many trials before convergence, and no direct interpretability of trained policies is provided. In order to achieve fast convergence and interpretability for the policy in RL, we propose a novel RL method for text-based games with a recent neuro-symbolic framework called Logical Neural Network, which can learn symbolic and interpretable rules in their differentiable network. The method is first to extract first-order logical facts from text observation and external word meaning network (ConceptNet), then train a policy in the network with directly interpretable logical operators. Our experimental results show RL training with the proposed method converges significantly faster than other state-of-the-art neuro-symbolic methods in a TextWorld benchmark.

Biomedical Concept Normalization by Leveraging Hypernyms
Cheng Yan et al.

Biomedical Concept Normalization (BCN) is widely used in biomedical text processing as a fundamental module. Owing to numerous surface variants of biomedical concepts, BCN still remains challenging and unsolved. In this paper, we exploit biomedical concept hypernyms to facilitate BCN. We propose Biomedical Concept Normalizer with Hypernyms (BCNH), a novel framework that adopts list-wise training to make use of both hypernyms and synonyms, and also employs norm constraint on the representation of hypernym-hyponym entity pairs. The experimental results show that BCNH outperforms the previous state-of-the-art model on the NCBI dataset.

Leveraging Capsule Routing to Associate Knowledge with Medical Literature Hierarchically
Xin Liu et al.

Integrating knowledge into text is a promising way to enrich text representation, especially in the medical field. However, undifferentiated knowledge not only confuses the text representation but also imports unexpected noises. In this paper, to alleviate this problem, we propose leveraging capsule routing to associate knowledge with medical literature hierarchically (called HiCapsRKL). Firstly, HiCapsRKL extracts two empirically designed text fragments from medical literature and encodes them into fragment representations respectively. Secondly, the capsule routing algorithm is applied to two fragment representations. Through the capsule computing and dynamic routing, each representation is processed into a new representation (denoted as caps-representation), and we integrate the caps-representations as information gain to associate knowledge with medical literature hierarchically. Finally, HiCapsRKL are validated on relevance prediction and medical literature retrieval test sets. The experimental results and analyses show that HiCapsRKLcan more accurately associate knowledge with medical literature than mainstream methods. In summary, HiCapsRKL can efficiently help selecting the most relevant knowledge to the medical literature, which may be an alternative attempt to improve knowledge-based text representation. Source code is released on GitHub.

Label-Enhanced Hierarchical Contextualized Representation for Sequential Metaphor Identification
Shuqun Li et al.

Recent metaphor identification approaches mainly consider the contextual text features within a sentence or introduce external linguistic features to the model. But they usually ignore the extra information that the data can provide, such as the contextual metaphor information and broader discourse information. In this paper, we propose a model augmented with hierarchical contextualized representation to extract more information from both sentence-level and discourse-level. At the sentence level, we leverage the metaphor information of words that except the target word in the sentence to strengthen the reasoning ability of our model via a novel label-enhanced contextualized representation. At the discourse level, the position-aware global memory network is adopted to learn long-range dependency among the same words within a discourse. Finally, our model combines the representations obtained from these two parts. The experiment results on two tasks of the VUA dataset show that our model outperforms every other state-of-the-art method that also does not use any external knowledge except what the pre-trained language model contains.

SpellBERT: A Lightweight Pretrained Model for Chinese Spelling Check
Tuo Ji, Hang Yan, and Xipeng Qiu

Chinese Spelling Check (CSC) is to detect and correct Chinese spelling errors. Many models utilize a predefined confusion set to learn a mapping between correct characters and its visually similar or phonetically similar misuses but the mapping may be out-of-domain. To that end, we propose SpellBERT, a pretrained model with graph-based extra features and independent on confusion set. To explicitly capture the two erroneous patterns, we employ a graph neural network to introduce radical and pinyin information as visual and phonetic features. For better fusing these features with character representations, we devise masked language model alike pre-training tasks. With this feature-rich pre-training, SpellBERT with only half size of BERT can show competitive performance and make a state-of-the-art result on the OCR dataset where most of the errors are not covered by the existing confusion set.

Automated Generation of Accurate & Fluent Medical X-ray Reports
Hoang Nguyen et al.

Our paper aims to automate the generation of medical reports from chest X-ray image inputs, a critical yet time-consuming task for radiologists. Existing medical report generation efforts emphasize producing human-readable reports, yet the generated text may not be well aligned to the clinical facts. Our generated medical reports, on the other hand, are fluent and, more importantly, clinically accurate. This is achieved by our fully differentiable and end-to-end paradigm that contains three complementary modules: taking the chest X-ray images and clinical history document of patients as inputs, our classification module produces an internal checklist of disease-related topics, referred to as enriched disease embedding; the embedding representation is then passed to our transformer-based generator, to produce the medical report; meanwhile, our generator also creates a weighted embedding representation, which is fed to our interpreter to ensure consistency with respect to disease-related topics. Empirical evaluations demonstrate very promising results achieved by our approach on commonly-used metrics concerning language fluency and clinical accuracy. Moreover, noticeable performance gains are consistently observed when additional input information is available, such as the clinical document and extra scans from different views.

Enhancing Document Ranking with Task-adaptive Training and Segmented Token Recovery Mechanism
Xingwu Sun et al.

In this paper, we propose a new ranking model DR-BERT, which improves the Document Retrieval (DR) task by a task-adaptive training process and a Segmented Token Recovery Mechanism (STRM). In the task-adaptive training, we first pre-train DR-BERT to be domain-adaptive and then make the two-phase fine-tuning. In the first-phase fine-tuning, the model learns query-document matching patterns regarding different query types in a pointwise way. Next, in the second-phase fine-tuning, the model learns document-level ranking features and ranks documents with regard to a given query in a listwise manner. Such pointwise plus listwise fine-tuning enables the model to minimize errors in the document ranking by incorporating ranking-specific supervisions. Meanwhile, the model derived from pointwise fine-tuning is also used to reduce noise in the training data of the listwise fine-tuning. On the other hand, we present STRM which can compute OOV word representation

and contextualization more precisely in BERT-based models. As an effective strategy in DR-BERT, STRM improves the matching perfromance of OOV words between a query and a document. Notably, our DR-BERT model keeps in the top three on the MS MARCO leaderboard since May 20, 2020.

Abstract, Rationale, Stance: A Joint Model for Scientific Claim Verification
Zhiwei Zhang et al.

Scientific claim verification can help the researchers to easily find the target scientific papers with the sentence evidence from a large corpus for the given claim. Some existing works propose pipeline models on the three tasks of abstract retrieval, rationale selection and stance prediction. Such works have the problems of error propagation among the modules in the pipeline and lack of sharing valuable information among modules. We thus propose an approach, named as ARSJoint, that jointly learns the modules for the three tasks with a machine reading comprehension framework by including claim information. In addition, we enhance the information exchanges and constraints among tasks by proposing a regularization term between the sentence attention scores of abstract retrieval and the estimated outputs of rational selection. The experimental results on the benchmark dataset SciFact show that our approach outperforms the existing works.

A Fine-Grained Domain Adaption Model for Joint Word Segmentation and POS Tagging
Peijie Jiang et al.

Domain adaption for word segmentation and POS tagging is a challenging problem for Chinese lexical processing. Self-training is one promising solution for it, which struggles to construct a set of high-quality pseudo training instances for the target domain. Previous work usually assumes a universal source-to-target adaption to collect such pseudo corpus, ignoring the different gaps from the target sentences to the source domain. In this work, we start from joint word segmentation and POS tagging, presenting a fine-grained domain adaption method to model the gaps accurately. We measure the gaps by one simple and intuitive metric, and adopt it to develop a pseudo target domain corpus based on fine-grained subdomains incrementally. A novel domain-mixed representation learning model is proposed accordingly to encode the multiple subdomains effectively. The whole process is performed progressively for both corpus construction and model training. Experimental results on a benchmark dataset show that our method can gain significant improvements over a vary of baselines. Extensive analyses are performed to show the advantages of our final domain adaption model as well.

Answering Open-Domain Questions of Varying Reasoning Steps from Text
Peng Qi et al.

We develop a unified system to answer directly from text open-domain questions that may require a varying number of retrieval steps. We employ a single multi-task transformer model to perform all the necessary subtasks—retrieving supporting facts, reranking them, and predicting the answer from all retrieved documents—in an iterative fashion. We avoid crucial assumptions of previous work that do not transfer well to real-world settings, including exploiting knowledge of the fixed number of retrieval steps required to answer each question or using structured metadata like knowledge bases or web links that have limited availability. Instead, we design a system that can answer open-domain questions on any text collection without prior knowledge of reasoning complexity. To emulate this setting, we construct a new benchmark, called BeerQA, by combining existing one- and two-step datasets with a new collection of 530 questions that require three Wikipedia pages to answer, unifying Wikipedia corpora versions in the process. We show that our model demonstrates competitive performance on both existing benchmarks and this new benchmark. We make the new benchmark available at https://beerqa.github.io/.

Adaptive Information Seeking for Open-Domain Question Answering
Yunchang Zhu et al.

Information seeking is an essential step for open-domain question answering to efficiently gather evidence from a large corpus. Recently, iterative approaches have been proven to be effective for complex questions, by recursively retrieving new evidence at each step. However, almost all existing iterative approaches use predefined strategies, either applying the same retrieval function multiple times or fixing the order of different retrieval functions, which cannot fulfill the diverse requirements of various questions. In this paper, we propose a novel adaptive information-seeking strategy for open-domain question answering, namely AISO. Specifically, the whole retrieval and answer process is modeled as a partially observed Markov decision process, where three types of retrieval operations (e.g., BM25, DPR, and hyperlink) and one answer operation are defined as actions. According to the learned policy, AISO could adaptively select a proper retrieval action to seek the missing evidence at each step, based on the collected evidence and the reformulated query, or directly output the answer when the evidence set is sufficient for the question. Experiments on SQuAD Open and HotpotQA fullwiki, which serve as single-hop and multi-hop open-domain QA benchmarks, show that AISO outperforms all baseline methods with predefined strategies in terms of both retrieval and answer evaluations.

Mapping probability word problems to executable representations
Simon Suster et al.

While solving math word problems automatically has received considerable attention in the NLP community, few works have addressed probability word problems specifically. In this paper, we employ and analyse various neural models for answering such word problems. In a two-step approach, the problem text is first mapped to a formal representation in a declarative language using a sequence-to-sequence model, and then the

resulting representation is executed using a probabilistic programming system to provide the answer. Our best performing model incorporates general-domain contextualised word representations that were finetuned using transfer learning on another in-domain dataset. We also apply end-to-end models to this task, which bring out the importance of the two-step approach in obtaining correct solutions to probability problems.

Enhancing Multiple-choice Machine Reading Comprehension by Punishing Illogical Interpretations
Yiming Ju et al.

Machine Reading Comprehension (MRC), which requires a machine to answer questions given the relevant documents, is an important way to test machines' ability to understand human language. Multiple-choice MRC is one of the most studied tasks in MRC due to the convenience of evaluation and the flexibility of answer format. Post-hoc interpretation aims to explain a trained model and reveal how the model arrives at the prediction. One of the most important interpretation forms is to attribute model decisions to input features. Based on post-hoc interpretation methods, we assess attributions of paragraphs in multiple-choice MRC and improve the model by punishing the illogical attributions. Our method can improve model performance without any external information and model structure change. Furthermore, we also analyze how and why such a self-training method works.

Large-Scale Relation Learning for Question Answering over Knowledge Bases with Pre-trained Language Models
Yuanmeng Yan et al.

The key challenge of question answering over knowledge bases (KBQA) is the inconsistency between the natural language questions and the reasoning paths in the knowledge base (KB). Recent graph-based KBQA methods are good at grasping the topological structure of the graph but often ignore the textual information carried by the nodes and edges. Meanwhile, pre-trained language models learn massive open-world knowledge from the large corpus, but it is in the natural language form and not structured. To bridge the gap between the natural language and the structured KB, we propose three relation learning tasks for BERT-based KBQA, including relation extraction, relation matching, and relation reasoning. By relation-augmented training, the model learns to align the natural language expressions to the relations in the KB as well as reason over the missing connections in the KB. Experiments on WebQSP show that our method consistently outperforms other baselines, especially when the KB is incomplete.

Phrase Retrieval Learns Passage Retrieval, Too
Jinhyuk Lee, Alexander Wettig, and Danqi Chen

Dense retrieval methods have shown great promise over sparse retrieval methods in a range of NLP problems. Among them, dense phrase retrieval—the most fine-grained retrieval unit—is appealing because phrases can be directly used as the output for question answering and slot filling tasks. In this work, we follow the intuition that retrieving phrases naturally entails retrieving larger text blocks and study whether phrase retrieval can serve as the basis for coarse-level retrieval including passages and documents. We first observe that a dense phrase-retrieval system, without any retraining, already achieves better passage retrieval accuracy (+3-5% in top-5 accuracy) compared to passage retrievers, which also helps achieve superior end-to-end QA performance with fewer passages. Then, we provide an interpretation for why phrase-level supervision helps learn better fine-grained entailment compared to passage-level supervision, and also show that phrase retrieval can be improved to achieve competitive performance in document-retrieval tasks such as entity linking and knowledge-grounded dialogue. Finally, we demonstrate how phrase filtering and vector quantization can reduce the size of our index by 4-10x, making dense phrase retrieval a practical and versatile solution in multi-granularity retrieval.

Neural Natural Logic Inference for Interpretable Question Answering
Jihao Shi et al.

Many open-domain question answering problems can be cast as a textual entailment task, where a question and candidate answers are concatenated to form hypotheses. A QA system then determines if the supporting knowledge bases, regarded as potential premises, entail the hypotheses. In this paper, we investigate a neural-symbolic QA approach that integrates natural logic reasoning within deep learning architectures, towards developing effective and yet explainable question answering models. The proposed model gradually bridges a hypothesis and candidate premises following natural logic inference steps to build proof paths. Entailment scores between the acquired intermediate hypotheses and candidate premises are measured to determine if a premise entails the hypothesis. As the natural logic reasoning process forms a tree-like, hierarchical structure, we embed hypotheses and premises in a Hyperbolic space rather than Euclidean space to acquire more precise representations. Empirically, our method outperforms prior work on answering multiple-choice science questions, achieving the best results on two publicly available datasets. The natural logic inference process inherently provides evidence to help explain the prediction process.

Smoothing Dialogue States for Open Conversational Machine Reading
Zhuosheng Zhang et al.

Conversational machine reading (CMR) requires machines to communicate with humans through multi-turn interactions between two salient dialogue states of decision making and question generation processes. In open CMR settings, as the more realistic scenario, the retrieved background knowledge would be noisy, which results in severe challenges in the information transmission. Existing studies commonly train independent or

pipeline systems for the two subtasks. However, those methods are trivial by using hard-label decisions to activate question generation, which eventually hinders the model performance. In this work, we propose an effective gating strategy by smoothing the two dialogue states in only one decoder and bridge decision making and question generation to provide a richer dialogue state reference. Experiments on the OR-ShARC dataset show the effectiveness of our method, which achieves new state-of-the-art results.

RockNER: A Simple Method to Create Adversarial Examples for Evaluating the Robustness of Named Entity Recognition Models
Bill Yuchen Lin et al.

To audit the robustness of named entity recognition (NER) models, we propose RockNER, a simple yet effective method to create natural adversarial examples. Specifically, at the entity level, we replace target entities with other entities of the same semantic class in Wikidata; at the context level, we use pre-trained language models (e.g., BERT) to generate word substitutions. Together, the two levels of at- tack produce natural adversarial examples that result in a shifted distribution from the training data on which our target models have been trained. We apply the proposed method to the OntoNotes dataset and create a new benchmark named OntoRock for evaluating the robustness of existing NER models via a systematic evaluation protocol. Our experiments and analysis reveal that even the best model has a significant performance drop, and these models seem to memorize in-domain entity patterns instead of reasoning from the context. Our work also studies the effects of a few simple data augmentation methods to improve the robustness of NER models.

Diagnosing the First-Order Logical Reasoning Ability Through LogicNLI
Jidong Tian et al.

Recently, language models (LMs) have achieved significant performance on many NLU tasks, which has spurred widespread interest for their possible applications in the scientific and social area. However, LMs have faced much criticism of whether they are truly capable of reasoning in NLU. In this work, we propose a diagnostic method for first-order logic (FOL) reasoning with a new proposed benchmark, LogicNLI. LogicNLI is an NLI-style dataset that effectively disentangles the target FOL reasoning from commonsense inference and can be used to diagnose LMs from four perspectives: accuracy, robustness, generalization, and interpretability. Experiments on BERT, RoBERTa, and XLNet, have uncovered the weaknesses of these LMs on FOL reasoning, which motivates future exploration to enhance the reasoning ability.

Constructing a Psychometric Testbed for Fair Natural Language Processing
Ahmed Abbasi et al.

Psychometric measures of ability, attitudes, perceptions, and beliefs are crucial for understanding user behavior in various contexts including health, security, e-commerce, and finance. Traditionally, psychometric dimensions have been measured and collected using survey-based methods. Inferring such constructs from user-generated text could allow timely, unobtrusive collection and analysis. In this paper we describe our efforts to construct a corpus for psychometric natural language processing (NLP) related to important dimensions such as trust, anxiety, numeracy, and literacy, in the health domain. We discuss our multi-step process to align user text with their survey-based response items and provide an overview of the resulting testbed which encompasses survey-based psychometric measures and accompanying user-generated text from 8,502 respondents. Our testbed also encompasses self-reported demographic information, including race, sex, age, income, and education - thereby affording opportunities for measuring bias and benchmarking fairness of text classification methods. We report preliminary results on use of the text to predict/categorize users' survey response labels - and on the fairness of these models. We also discuss the important implications of our work and resulting testbed for future NLP research on psychometrics and fairness.

COUGH: A Challenge Dataset and Models for COVID-19 FAQ Retrieval
Xinliang Frederick Zhang et al.

We present a large, challenging dataset, COUGH, for COVID-19 FAQ retrieval. Similar to a standard FAQ dataset, COUGH consists of three parts: FAQ Bank, Query Bank and Relevance Set. The FAQ Bank contains ~16K FAQ items scraped from 55 credible websites (e.g., CDC and WHO). For evaluation, we introduce Query Bank and Relevance Set, where the former contains 1,236 human-paraphrased queries while the latter contains ~32 human-annotated FAQ items for each query. We analyze COUGH by testing different FAQ retrieval models built on top of BM25 and BERT, among which the best model achieves 48.8 under P5, indicating a great challenge presented by COUGH and encouraging future research for further improvement. Our COUGH dataset is available at https://github.com/sunlab-osu/covid-faq.

Chinese WPLC: A Chinese Dataset for Evaluating Pretrained Language Models on Word Prediction Given Long-Range Context
Huibin Ge et al.

This paper presents a Chinese dataset for evaluating pretrained language models on Word Prediction given Long-term Context (Chinese WPLC). We propose both automatic and manual selection strategies tailored to Chinese to guarantee that target words in passages collected from over 69K novels can only be predicted with long-term context beyond the scope of sentences containing the target words. Dataset analysis reveals that the types of target words range from common nouns to Chinese 4-character idioms. We also observe that linguistic relations between target words and long-range context exhibit diversity, including lexical match, synonym, summary and reasoning. Experiment results show that the Chinese pretrained language model PanGu-α is 45

points behind human in terms of top-1 word prediction accuracy, indicating that Chinese WPLC is a challenging dataset. The dataset is publicly available at https://git.openi.org.cn/PCL-Platform.Intelligence/Chinese $_{W}PLC$.

WinoLogic: A Zero-Shot Logic-based Diagnostic Dataset for Winograd Schema Challenge
Weinan He et al.

The recent success of neural language models (NLMs) on the Winograd Schema Challenge has called for further investigation of the commonsense reasoning ability of these models. Previous diagnostic datasets rely on crowd-sourcing which fails to provide coherent commonsense crucial for solving WSC problems. To better evaluate NLMs, we propose a logic-based framework that focuses on high-quality commonsense knowledge. Specifically, we identify and collect formal knowledge formulas verified by theorem provers and translate such formulas into natural language sentences. Based on these true knowledge sentences, adversarial false ones are generated. We propose a new dataset named WinoLogic with these sentences. Given a problem in WinoLogic, NLMs need to decide whether the plausible knowledge sentences could correctly solve the corresponding WSC problems in a zero-shot setting. We also ask human annotators to validate WinoLogic to ensure it is human-agreeable. Experiments show that NLMs still struggle to comprehend commonsense knowledge as humans do, indicating that their reasoning ability could have been overestimated.

Pseudo Zero Pronoun Resolution Improves Zero Anaphora Resolution
Ryuto Konno et al.

Masked language models (MLMs) have contributed to drastic performance improvements with regard to zero anaphora resolution (ZAR). To further improve this approach, in this study, we made two proposals. The first is a new pretraining task that trains MLMs on anaphoric relations with explicit supervision, and the second proposal is a new finetuning method that remedies a notorious issue, the pretrain-finetune discrepancy. Our experiments on Japanese ZAR demonstrated that our two proposals boost the state-of-the-art performance, and our detailed analysis provides new insights on the remaining challenges.

Aligning Cross-lingual Sentence Representations with Dual Momentum Contrast
Liang Wang, Wei Zhao, and Jingming Liu

In this paper, we propose to align sentence representations from different languages into a unified embedding space, where semantic similarities (both cross-lingual and monolingual) can be computed with a simple dot product. Pre-trained language models are fine-tuned with the translation ranking task. Existing work (Feng et al., 2020) uses sentences within the same batch as negatives, which can suffer from the issue of easy negatives. We adapt MoCo (He et al., 2020) to further improve the quality of alignment. As the experimental results show, the sentence representations produced by our model achieve the new state-of-the-art on several tasks, including Tatoeba en-zh similarity search (Artetxe andSchwenk, 2019b), BUCC en-zh bitext mining, and semantic textual similarity on 7 datasets.

Total Recall: a Customized Continual Learning Method for Neural Semantic Parsers
Zhuang Li, Lizhen Qu, and Gholamreza Haffari

This paper investigates continual learning for semantic parsing. In this setting, a neural semantic parser learns tasks sequentially without accessing full training data from previous tasks. Direct application of the SOTA continual learning algorithms to this problem fails to achieve comparable performance with re-training models with all seen tasks because they have not considered the special properties of structured outputs yielded by semantic parsers. Therefore, we propose TotalRecall, a continual learning method designed for neural semantic parsers from two aspects: i) a sampling method for memory replay that diversifies logical form templates and balances distributions of parse actions in a memory; ii) a two-stage training method that significantly improves generalization capability of the parsers across tasks. We conduct extensive experiments to study the research problems involved in continual semantic parsing and demonstrate that a neural semantic parser trained with TotalRecall achieves superior performance than the one trained directly with the SOTA continual learning algorithms and achieve a 3-6 times speedup compared to re-training from scratch.

Exophoric Pronoun Resolution in Dialogues with Topic Regularization
Xintong Yu et al.

Resolving pronouns to their referents has long been studied as a fundamental natural language understanding problem. Previous works on pronoun coreference resolution (PCR) mostly focus on resolving pronouns to mentions in text while ignoring the exophoric scenario. Exophoric pronouns are common in daily communications, where speakers may directly use pronouns to refer to some objects present in the environment without introducing the objects first. Although such objects are not mentioned in the dialogue text, they can often be disambiguated by the general topics of the dialogue. Motivated by this, we propose to jointly leverage the local context and global topics of dialogues to solve the out-of-text PCR problem. Extensive experiments demonstrate the effectiveness of adding topic regularization for resolving exophoric pronouns.

Context-Aware Interaction Network for Question Matching
Zhe Hu et al.

Impressive milestones have been achieved in text matching by adopting a cross-attention mechanism to capture pertinent semantic connections between two sentence representations. However, regular cross-attention focuses on word-level links between the two input sequences, neglecting the importance of contextual infor-

mation. We propose a context-aware interaction network (COIN) to properly align two sequences and infer their semantic relationship. Specifically, each interaction block includes (1) a context-aware cross-attention mechanism to effectively integrate contextual information when aligning two sequences, and (2) a gate fusion layer to flexibly interpolate aligned representations. We apply multiple stacked interaction blocks to produce alignments at different levels and gradually refine the attention results. Experiments on two question matching datasets and detailed analyses demonstrate the effectiveness of our model.

TEMP: Taxonomy Expansion with Dynamic Margin Loss through Taxonomy-Paths

Zichen Liu et al.

As an essential form of knowledge representation, taxonomies are widely used in various downstream natural language processing tasks. However, with the continuously rising of new concepts, many existing taxonomies are unable to maintain coverage by manual expansion. In this paper, we propose TEMP, a self-supervised taxonomy expansion method, which predicts the position of new concepts by ranking the generated taxonomy-paths. For the first time, TEMP employs pre-trained contextual encoders in taxonomy construction and hypernym detection problems. Experiments prove that pre-trained contextual embeddings are able to capture hypernym-hyponym relations. To learn more detailed differences between taxonomy-paths, we train the model with dynamic margin loss by a novel dynamic margin function. Extensive evaluations exhibit that TEMP outperforms prior state-of-the-art taxonomy expansion approaches by 14.3% in accuracy and 15.8% in mean reciprocal rank on three public benchmarks.

A Graph-Based Neural Model for End-to-End Frame Semantic Parsing

ZhiChao Lin, Yueheng Sun, and Meishan Zhang

Frame semantic parsing is a semantic analysis task based on FrameNet which has received great attention recently. The task usually involves three subtasks sequentially: (1) target identification, (2) frame classification and (3) semantic role labeling. The three subtasks are closely related while previous studies model them individually, which ignores their intern connections and meanwhile induces error propagation problem. In this work, we propose an end-to-end neural model to tackle the task jointly. Concretely, we exploit a graph-based method, regarding frame semantic parsing as a graph construction problem. All predicates and roles are treated as graph nodes, and their relations are taken as graph edges. Experiment results on two benchmark datasets of frame semantic parsing show that our method is highly competitive, resulting in better performance than pipeline models.

Virtual Data Augmentation: A Robust and General Framework for Fine-tuning Pre-trained Models

Kun Zhou et al.

Recent works have shown that powerful pre-trained language models (PLM) can be fooled by small perturbations or intentional attacks. To solve this issue, various data augmentation techniques are proposed to improve the robustness of PLMs. However, it is still challenging to augment semantically relevant examples with sufficient diversity. In this work, we present Virtual Data Augmentation (VDA), a general framework for robustly fine-tuning PLMs. Based on the original token embeddings, we construct a multinomial mixture for augmenting virtual data embeddings, where a masked language model guarantees the semantic relevance and the Gaussian noise provides the augmentation diversity. Furthermore, a regularized training strategy is proposed to balance the two aspects. Extensive experiments on six datasets show that our approach is able to improve the robustness of PLMs and alleviate the performance degradation under adversarial attacks. Our codes and data are publicly available at https://github.com/RUCAIBox/VDA.

CATE: A Contrastive Pre-trained Model for Metaphor Detection with Semi-supervised Learning

Zhenxi Lin et al.

Metaphors are ubiquitous in natural language, and detecting them requires contextual reasoning about whether a semantic incongruence actually exists. Most existing work addresses this problem using pre-trained contextualized models. Despite their success, these models require a large amount of labeled data and are not linguistically-based. In this paper, we proposed a ContrAstive pre-Trained modEl (CATE) for metaphor detection with semi-supervised learning. Our model first uses a pre-trained model to obtain a contextual representation of target words and employs a contrastive objective to promote an increased distance between target words' literal and metaphorical senses based on linguistic theories. Furthermore, we propose a simple strategy to collect large-scale candidate instances from the general corpus and generalize the model via self-training. Extensive experiments show that CATE achieves better performance against state-of-the-art baselines on several benchmark datasets.

To be Closer: Learning to Link up Aspects with Opinions

Yuxiang Zhou et al.

Dependency parse trees are helpful for discovering the opinion words in aspect-based sentiment analysis (ABSA)~[**huang2019syntax**]. However, the trees obtained from off-the-shelf dependency parsers are static, and could be sub-optimal in ABSA. This is because the syntactic trees are not designed for capturing the interactions between opinion words and aspect words. In this work, we aim to shorten the distance between aspects and corresponding opinion words by learning an aspect-centric tree structure. The aspect and opinion words are expected to be closer along such tree structure compared to the standard dependency parse tree. The

learning process allows the tree structure to adaptively correlate the aspect and opinion words, enabling us to better identify the polarity in the ABSA task. We conduct experiments on five aspect-based sentiment datasets, and the proposed model significantly outperforms recent strong baselines. Furthermore, our thorough analysis demonstrates the average distance between aspect and opinion words are shortened by at least 19% on the standard SemEval Restaurant14~Pontiki2014SemEval2014T4 dataset.

Seeking Common but Distinguishing Difference, A Joint Aspect-based Sentiment Analysis Model
Hongjiang Jing et al.

Aspect-based sentiment analysis (ABSA) task consists of three typical subtasks: aspect term extraction, opinion term extraction, and sentiment polarity classification. These three subtasks are usually performed jointly to save resources and reduce the error propagation in the pipeline. However, most of the existing joint models only focus on the benefits of encoder sharing between subtasks but ignore the difference. Therefore, we propose a joint ABSA model, which not only enjoys the benefits of encoder sharing but also focuses on the difference to improve the effectiveness of the model. In detail, we introduce a dual-encoder design, in which a pair encoder especially focuses on candidate aspect-opinion pair classification, and the original encoder keeps attention on sequence labeling. Empirical results show that our proposed model shows robustness and significantly outperforms the previous state-of-the-art on four benchmark datasets.

Argument Pair Extraction with Mutual Guidance and Inter-sentence Relation Graph
Jianzhu Bao et al.

Argument pair extraction (APE) aims to extract interactive argument pairs from two passages of a discussion. Previous work studied this task in the context of peer review and rebuttal, and decomposed it into a sequence labeling task and a sentence relation classification task. However, despite the promising performance, such an approach obtains the argument pairs implicitly by the two decomposed tasks, lacking explicitly modeling of the argument-level interactions between argument pairs. In this paper, we tackle the APE task by a mutual guidance framework, which could utilize the information of an argument in one passage to guide the identification of arguments that can form pairs with it in another passage. In this manner, two passages can mutually guide each other in the process of APE. Furthermore, we propose an inter-sentence relation graph to effectively model the inter-relations between two sentences and thus facilitates the extraction of argument pairs. Our proposed method can better represent the holistic argument-level semantics and thus explicitly capture the complex correlations between argument pairs. Experimental results show that our approach significantly outperforms the current state-of-the-art model.

Emotion Inference in Multi-Turn Conversations with Addressee-Aware Module and Ensemble Strategy
Dayu Li et al.

Emotion inference in multi-turn conversations aims to predict the participant's emotion in the next upcoming turn without knowing the participant's response yet, and is a necessary step for applications such as dialogue planning. However, it is a severe challenge to perceive and reason about the future feelings of participants, due to the lack of utterance information from the future. Moreover, it is crucial for emotion inference to capture the characteristics of emotional propagation in conversations, such as persistence and contagiousness. In this study, we focus on investigating the task of emotion inference in multi-turn conversations by modeling the propagation of emotional states among participants in the conversation history, and propose an addressee-aware module to automatically learn whether the participant keeps the historical emotional state or is affected by others in the next upcoming turn. In addition, we propose an ensemble strategy to further enhance the model performance. Empirical studies on three different benchmark conversation datasets demonstrate the effectiveness of the proposed model over several strong baselines.

Improving Federated Learning for Aspect-based Sentiment Analysis via Topic Memories
Han Qin et al.

Aspect-based sentiment analysis (ABSA) predicts the sentiment polarity towards a particular aspect term in a sentence, which is an important task in real-world applications. To perform ABSA, the trained model is required to have a good understanding of the contextual information, especially the particular patterns that suggest the sentiment polarity. However, these patterns typically vary in different sentences, especially when the sentences come from different sources (domains), which makes ABSA still very challenging. Although combining labeled data across different sources (domains) is a promising solution to address the challenge, in practical applications, these labeled data are usually stored at different locations and might be inaccessible to each other due to privacy or legal concerns (e.g., the data are owned by different companies). To address this issue and make the best use of all labeled data, we propose a novel ABSA model with federated learning (FL) adopted to overcome the data isolation limitations and incorporate topic memory (TM) proposed to take the cases of data from diverse sources (domains) into consideration. Particularly, TM aims to identify different isolated data sources due to data inaccessibility by providing useful categorical information for localized predictions. Experimental results on a simulated environment for FL with three nodes demonstrate the effectiveness of our approach, where TM-FL outperforms different baselines including some well-designed FL frameworks.

Comparative Opinion Quintuple Extraction from Product Reviews
Ziheng Liu, Rui Xia, and Jianfei Yu

As an important task in opinion mining, comparative opinion mining aims to identify comparative sentences from product reviews, extract the comparative elements, and obtain the corresponding comparative opinion tuples. However, most previous studies simply regarded comparative tuple extraction as comparative element extraction, but ignored the fact that many comparative sentences may contain multiple comparisons. The comparative opinion tuples defined in these studies also failed to explicitly provide comparative preferences. To address these limitations, in this work we first introduce a new Comparative Opinion Quintuple Extraction (COQE) task, to identify comparative sentences from product reviews and extract all comparative opinion quintuples (Subject, Object, Comparative Aspect, Comparative Opinion, Comparative Preference). Secondly, based on the existing comparative opinion mining corpora, we make supplementary annotations and construct three datasets for the COQE task. Finally, we benchmark the COQE task by proposing a new BERT-based multi-stage approach as well as three baseline systems extended from previous methods. %The new approach significantly outperforms three baseline systems on three datasets and represents a strong benchmark for COQE. Experimental results show that the new approach significantly outperforms three baseline systems on three datasets for the COQE task.

CTAL: Pre-training Cross-modal Transformer for Audio-and-Language Representations

Hang Li et al.

Existing audio-language task-specific predictive approaches focus on building complicated late-fusion mechanisms. However, these models are facing challenges of overfitting with limited labels and low model generalization abilities. In this paper, we present a Cross-modal Transformer for Audio-and-Language, i.e., CTAL, which aims to learn the intra-modality and inter-modality connections between audio and language through two proxy tasks on a large amount of audio-and-language pairs: masked language modeling and masked cross-modal acoustic modeling. After fine-tuning our pre-trained model on multiple downstream audio-and-language tasks, we observe significant improvements across various tasks, such as, emotion classification, sentiment analysis, and speaker verification. On this basis, we further propose a specially-designed fusion mechanism that can be used in fine-tuning phase, which allows our pre-trained model to achieve better performance. Lastly, we demonstrate detailed ablation studies to prove that both our novel cross-modality fusion component and audio-language pre-training methods significantly contribute to the promising results. The code and pre-trained models are available at https://github.com/tal-ai/CTAL_EMNLP2021.

Relation-aware Video Reading Comprehension for Temporal Language Grounding

Jialin Gao et al.

Temporal language grounding in videos aims to localize the temporal span relevant to the given query sentence. Previous methods treat it either as a boundary regression task or a span extraction task. This paper will formulate temporal language grounding into video reading comprehension and propose a Relation-aware Network (RaNet) to address it. This framework aims to select a video moment choice from the predefined answer set with the aid of coarse-and-fine choice-query interaction and choice-choice relation construction. A choice-query interactor is proposed to match the visual and textual information simultaneously in sentence-moment and token-moment levels, leading to a coarse-and-fine cross-modal interaction. Moreover, a novel multi-choice relation constructor is introduced by leveraging graph convolution to capture the dependencies among video moment choices for the best choice selection. Extensive experiments on ActivityNet-Captions, TACoS, and Charades-STA demonstrate the effectiveness of our solution. Codes will be available at https://github.com/Huntersxsx/RaNet.

Mutual-Learning Improves End-to-End Speech Translation

Jiawei Zhao et al.

A currently popular research area in end-to-end speech translation is the use of knowledge distillation from a machine translation (MT) task to improve the speech translation (ST) task. However, such scenario obviously only allows one way transfer, which is limited by the performance of the teacher model. Therefore, We hypothesis that the knowledge distillation-based approaches are sub-optimal. In this paper, we propose an alternative—a trainable mutual-learning scenario, where the MT and the ST models are collaboratively trained and are considered as peers, rather than teacher/student. This allows us to improve the performance of end-to-end ST more effectively than with a teacher-student paradigm. As a side benefit, performance of the MT model also improves. Experimental results show that in our mutual-learning scenario, models can effectively utilise the auxiliary information from peer models and achieve compelling results on Must-C dataset.

Vision Guided Generative Pre-trained Language Models for Multimodal Abstractive Summarization

Tiezheng Yu et al.

Multimodal abstractive summarization (MAS) models that summarize videos (vision modality) and their corresponding transcripts (text modality) are able to extract the essential information from massive multimodal data on the Internet. Recently, large-scale generative pre-trained language models (GPLMs) have been shown to be effective in text generation tasks. However, existing MAS models cannot leverage GPLMs' powerful generation ability. To fill this research gap, we aim to study two research questions: 1) how to inject visual information into GPLMs without hurting their generation ability; and 2) where is the optimal place in GPLMs to inject the visual information? In this paper, we present a simple yet effective method to construct vision guided (VG) GPLMs for the MAS task using attention-based add-on layers to incorporate visual information while maintaining their original text generation ability. Results show that our best model significantly surpasses the prior state-of-the-art model by 5.7 ROUGE-1, 5.3 ROUGE-2, and 5.1 ROUGE-L scores on the

How2 dataset, and our vision guidance method contributes 83.6% of the overall improvement. Furthermore, we conduct thorough ablation studies to analyze the effectiveness of various modality fusion methods and fusion locations.

Natural Language Video Localization with Learnable Moment Proposals
Shaoning Xiao et al.

Given an untrimmed video and a natural language query, Natural Language Video Localization (NLVL) aims to identify the video moment described by query. To address this task, existing methods can be roughly grouped into two groups: 1) propose-and-rank models first define a set of hand-designed moment candidates and then find out the best-matching one. 2) proposal-free models directly predict two temporal boundaries of the referential moment from frames. Currently, almost all the propose-and-rank methods have inferior performance than proposal-free counterparts. In this paper, we argue that the performance of propose-and-rank models are underestimated due to the predefined manners: 1) Hand-designed rules are hard to guarantee the complete coverage of targeted segments. 2) Densely sampled candidate moments cause redundant computation and degrade the performance of ranking process. To this end, we propose a novel model termed LPNet (Learnable Proposal Network for NLVL) with a fixed set of learnable moment proposals. The position and length of these proposals are dynamically adjusted during training process. Moreover, a boundary-aware loss has been proposed to leverage frame-level information and further improve performance. Extensive ablations on two challenging NLVL benchmarks have demonstrated the effectiveness of LPNet over existing state-of-the-art methods.

Language-Aligned Waypoint (LAW) Supervision for Vision-and-Language Navigation in Continuous Environments
Sonia Raychaudhuri et al.

In the Vision-and-Language Navigation (VLN) task an embodied agent navigates a 3D environment, following natural language instructions. A challenge in this task is how to handle 'off the path' scenarios where an agent veers from a reference path. Prior work supervises the agent with actions based on the shortest path from the agent's location to the goal, but such goal-oriented supervision is often not in alignment with the instruction. Furthermore, the evaluation metrics employed by prior work do not measure how much of a language instruction the agent is able to follow. In this work, we propose a simple and effective language-aligned supervision scheme, and a new metric that measures the number of sub-instructions the agent has completed during navigation.

How to leverage the multimodal EHR data for better medical prediction?
Bo Yang and Lijun Wu

Healthcare is becoming a more and more important research topic recently. With the growing data in the healthcare domain, it offers a great opportunity for deep learning to improve the quality of service and reduce costs. However, the complexity of electronic health records~(EHR) data is a challenge for the application of deep learning. Specifically, the data produced in the hospital admissions are monitored by the EHR system, which includes structured data like daily body temperature and unstructured data like free text and laboratory measurements. Although there are some preprocessing frameworks proposed for specific EHR data, the clinical notes that contain significant clinical value are beyond the realm of their consideration. Besides, whether these different data from various views are all beneficial to the medical tasks and how to best utilize these data remain unclear. Therefore, in this paper, we first extract the accompanying clinical notes from EHR and propose a method to integrate these data, we also comprehensively study the different models and the data leverage methods for better medical task prediction performance. The results on two prediction tasks show that our fused model with different data outperforms the state-of-the-art method without clinical notes, which illustrates the importance of our fusion method and the clinical note features.

Considering Nested Tree Structure in Sentence Extractive Summarization with Pre-trained Transformer
Jingun Kwon et al.

Sentence extractive summarization shortens a document by selecting sentences for a summary while preserving its important contents. However, constructing a coherent and informative summary is difficult using a pre-trained BERT-based encoder since it is not explicitly trained for representing the information of sentences in a document. We propose a nested tree-based extractive summarization model on RoBERTa (NeRoBERTa), where nested tree structures consist of syntactic and discourse trees in a given document. Experimental results on the CNN/DailyMail dataset showed that NeRoBERTa outperforms baseline models in ROUGE. Human evaluation results also showed that NeRoBERTa achieves significantly better scores than the baselines in terms of coherence and yields comparable scores to the state-of-the-art models.

Frame Semantic-Enhanced Sentence Modeling for Sentence-level Extractive Text Summarization
Yong Guan et al.

Sentence-level extractive text summarization aims to select important sentences from a given document. However, it is very challenging to model the importance of sentences. In this paper, we propose a novel Frame Semantic-Enhanced Sentence Modeling for Extractive Summarization, which leverages Frame semantics to model sentences from both intra-sentence level and inter-sentence level, facilitating the text summarization task. In particular, intra-sentence level semantics leverage Frames and Frame Elements to model internal se-

mantic structure within a sentence, while inter-sentence level semantics leverage Frame-to-Frame relations to model relationships among sentences. Extensive experiments on two benchmark corpus CNN/DM and NYT demonstrate that our model outperforms six state-of-the-art methods significantly.

CAST: Enhancing Code Summarization with Hierarchical Splitting and Reconstruction of Abstract Syntax Trees

Ensheng Shi et al.

Code summarization aims to generate concise natural language descriptions of source code, which can help improve program comprehension and maintenance. Recent studies show that syntactic and structural information extracted from abstract syntax trees (ASTs) is conducive to summary generation. However, existing approaches fail to fully capture the rich information in ASTs because of the large size/depth of ASTs. In this paper, we propose a novel model CAST that hierarchically splits and reconstructs ASTs. First, we hierarchically split a large AST into a set of subtrees and utilize a recursive neural network to encode the subtrees. Then, we aggregate the embeddings of subtrees by reconstructing the split ASTs to get the representation of the complete AST. Finally, AST representation, together with source code embedding obtained by a vanilla code token encoder, is used for code summarization. Extensive experiments, including the ablation study and the human evaluation, on benchmarks have demonstrated the power of CAST. To facilitate reproducibility, our code and data are available at https://github.com/DeepSoftwareAnalytics/CAST.

SgSum:Transforming Multi-document Summarization into Sub-graph Selection

Moye Chen et al.

Most of existing extractive multi-document summarization (MDS) methods score each sentence individually and extract salient sentences one by one to compose a summary, which have two main drawbacks: (1) neglecting both the intra and cross-document relations between sentences; (2) neglecting the coherence and conciseness of the whole summary. In this paper, we propose a novel MDS framework (SgSum) to formulate the MDS task as a sub-graph selection problem, in which source documents are regarded as a relation graph of sentences (e.g., similarity graph or discourse graph) and the candidate summaries are its sub-graphs. Instead of selecting salient sentences, SgSum selects a salient sub-graph from the relation graph as the summary. Comparing with traditional methods, our method has two main advantages: (1) the relations between sentences are captured by modeling both the graph structure of the whole document set and the candidate sub-graphs; (2) directly outputs an integrate summary in the form of sub-graph which is more informative and coherent. Extensive experiments on MultiNews and DUC datasets show that our proposed method brings substantial improvements over several strong baselines. Human evaluation results also demonstrate that our model can produce significantly more coherent and informative summaries compared with traditional MDS methods. Moreover, the proposed architecture has strong transfer ability from single to multi-document input, which can reduce the resource bottleneck in MDS tasks.

Event Graph based Sentence Fusion

Ruifeng Yuan, zili Wang, and Wenjie Li

Sentence fusion is a conditional generation task that merges several related sentences into a coherent one, which can be deemed as a summary sentence. The importance of sentence fusion has long been recognized by communities in natural language generation, especially in text summarization. It remains challenging for a state-of-the-art neural abstractive summarization model to generate a well-integrated summary sentence. In this paper, we explore the effective sentence fusion method in the context of text summarization. We propose to build an event graph from the input sentences to effectively capture and organize related events in a structured way and use the constructed event graph to guide sentence fusion. In addition to make use of the attention over the content of sentences and graph nodes, we further develop a graph flow attention mechanism to control the fusion process via the graph structure. When evaluated on sentence fusion data built from two summarization datasets, CNN/DaliyMail and Multi-News, our model shows to achieve state-of-the-art performance in terms of Rouge and other metrics like fusion rate and faithfulness.

Transformer-based Lexically Constrained Headline Generation

Kosuke Yamada et al.

This paper explores a variant of automatic headline generation methods, where a generated headline is required to include a given phrase such as a company or a product name. Previous methods using Transformer-based models generate a headline including a given phrase by providing the encoder with additional information corresponding to the given phrase. However, these methods cannot always include the phrase in the generated headline. Inspired by previous RNN-based methods generating token sequences in backward and forward directions from the given phrase, we propose a simple Transformer-based method that guarantees to include the given phrase in the high-quality generated headline. We also consider a new headline generation strategy that takes advantage of the controllable generation order of Transformer. Our experiments with the Japanese News Corpus demonstrate that our methods, which are guaranteed to include the phrase in the generated headline, achieve ROUGE scores comparable to previous Transformer-based methods. We also show that our generation strategy performs better than previous strategies.

Learn to Copy from the Copying History: Correlational Copy Network for Abstractive Summarization

Haoran Li et al.

The copying mechanism has had considerable success in abstractive summarization, facilitating models to directly copy words from the input text to the output summary. Existing works mostly employ encoder-decoder attention, which applies copying at each time step independently of the former ones. However, this may sometimes lead to incomplete copying. In this paper, we propose a novel copying scheme named Correlational Copying Network (CoCoNet) that enhances the standard copying mechanism by keeping track of the copying history. It thereby takes advantage of prior copying distributions and, at each time step, explicitly encourages the model to copy the input word that is relevant to the previously copied one. In addition, we strengthen CoCoNet through pre-training with suitable corpora that simulate the copying behaviors. Experimental results show that CoCoNet can copy more accurately and achieves new state-of-the-art performances on summarization benchmarks, including CNN/DailyMail for news summarization and SAMSum for dialogue summarization. The code and checkpoint will be publicly available.

Gradient-Based Adversarial Factual Consistency Evaluation for Abstractive Summarization

Zhiyuan Zeng et al.

Neural abstractive summarization systems have gained significant progress in recent years. However, abstractive summarization often produce inconsisitent statements or false facts. How to automatically generate highly abstract yet factually correct summaries? In this paper, we proposed an efficient weak-supervised adversarial data augmentation approach to form the factual consistency dataset. Based on the artificial dataset, we train an evaluation model that can not only make accurate and robust factual consistency discrimination but is also capable of making interpretable factual errors tracing by backpropagated gradient distribution on token embeddings. Experiments and analysis conduct on public annotated summarization and factual consistency datasets demonstrate our approach effective and reasonable.

Word Reordering for Zero-shot Cross-lingual Structured Prediction

Tao Ji et al.

Adapting word order from one language to another is a key problem in cross-lingual structured prediction. Current sentence encoders (e.g., RNN, Transformer with position embeddings) are usually word order sensitive. Even with uniform word form representations (MUSE, mBERT), word order discrepancies may hurt the adaptation of models. In this paper, we build structured prediction models with bag-of-words inputs, and introduce a new reordering module to organizing words following the source language order, which learns task-specific reordering strategies from a general-purpose order predictor model. Experiments on zero-shot cross-lingual dependency parsing, POS tagging, and morphological tagging show that our model can significantly improve target language performances, especially for languages that are distant from the source language.

A Unified Encoding of Structures in Transition Systems

Tao Ji et al.

Transition systems usually contain various dynamic structures (e.g., stacks, buffers). An ideal transition-based model should encode these structures completely and efficiently. Previous works relying on templates or neural network structures either only encode partial structure information or suffer from computation efficiency. In this paper, we propose a novel attention-based encoder unifying representation of all structures in a transition system. Specifically, we separate two views of items on structures, namely structure-invariant view and structure-dependent view. With the help of parallel-friendly attention network, we are able to encoding transition states with O(1) additional complexity (with respect to basic feature extractors). Experiments on the PTB and UD show that our proposed method significantly improves the test speed and achieves the best transition-based model, and is comparable to state-of-the-art methods.

Rumor Detection on Twitter with Claim-Guided Hierarchical Graph Attention Networks

Hongzhan Lin et al.

Rumors are rampant in the era of social media. Conversation structures provide valuable clues to differentiate between real and fake claims. However, existing rumor detection methods are either limited to the strict relation of user responses or oversimplify the conversation structure. In this study, to substantially reinforces the interaction of user opinions while alleviating the negative impact imposed by irrelevant posts, we first represent the conversation thread as an undirected interaction graph. We then present a Claim-guided Hierarchical Graph Attention Network for rumor classification, which enhances the representation learning for responsive posts considering the entire social contexts and attends over the posts that can semantically infer the target claim. Extensive experiments on three Twitter datasets demonstrate that our rumor detection method achieves much better performance than state-of-the-art methods and exhibits a superior capacity for detecting rumors at early stages.

Learning Bill Similarity with Annotated and Augmented Corpora of Bills

Jiseon Kim et al.

Bill writing is a critical element of representative democracy. However, it is often overlooked that most legislative bills are derived, or even directly copied, from other bills. Despite the significance of bill-to-bill linkages for understanding the legislative process, existing approaches fail to address semantic similarities across bills, let alone reordering or paraphrasing which are prevalent in legal document writing. In this paper, we overcome these limitations by proposing a 5-class classification task that closely reflects the nature of the bill generation process. In doing so, we construct a human-labeled dataset of 4,721 bill-to-bill relationships at

the subsection-level and release this annotated dataset to the research community. To augment the dataset, we generate synthetic data with varying degrees of similarity, mimicking the complex bill writing process. We use BERT variants and apply multi-stage training, sequentially fine-tuning our models with synthetic and human-labeled datasets. We find that the predictive performance significantly improves when training with both human-labeled and synthetic data. Finally, we apply our trained model to infer section- and bill-level similarities. Our analysis shows that the proposed methodology successfully captures the similarities across legal documents at various levels of aggregation.

CRFR: Improving Conversational Recommender Systems via Flexible Fragments Reasoning on Knowledge Graphs

Jinfeng Zhou et al.

Although paths of user interests shift in knowledge graphs (KGs) can benefit conversational recommender systems (CRS), explicit reasoning on KGs has not been well considered in CRS, due to the complex of high-order and incomplete paths. We propose CRFR, which effectively does explicit multi-hop reasoning on KGs with a conversational context-based reinforcement learning model. Considering the incompleteness of KGs, instead of learning single complete reasoning path, CRFR flexibly learns multiple reasoning fragments which are likely contained in the complete paths of interests shift. A fragments-aware unified model is then designed to fuse the fragments information from item-oriented and concept-oriented KGs to enhance the CRS response with entities and words from the fragments. Extensive experiments demonstrate CRFR's SOTA performance on recommendation, conversation and conversation interpretability.

Graph Based Network with Contextualized Representations of Turns in Dialogue

Bongseok Lee and Yong Suk Choi

Dialogue-based relation extraction (RE) aims to extract relation(s) between two arguments that appear in a dialogue. Because dialogues have the characteristics of high personal pronoun occurrences and low information density, and since most relational facts in dialogues are not supported by any single sentence, dialogue-based relation extraction requires a comprehensive understanding of dialogue. In this paper, we propose the TUrn COntext awaRE Graph Convolutional Network (TUCORE-GCN) modeled by paying attention to the way people understand dialogues. In addition, we propose a novel approach which treats the task of emotion recognition in conversations (ERC) as a dialogue-based RE. Experiments on a dialogue-based RE dataset and three ERC datasets demonstrate that our model is very effective in various dialogue-based natural language understanding tasks. In these experiments, TUCORE-GCN outperforms the state-of-the-art models on most of the benchmark datasets. Our code is available at https://github.com/BlackNoodle/TUCORE-GCN.

Detecting Speaker Personas from Conversational Texts

Jia-Chen Gu et al.

Personas are useful for dialogue response prediction. However, the personas used in current studies are predefined and hard to obtain before a conversation. To tackle this issue, we study a new task, named Speaker Persona Detection (SPD), which aims to detect speaker personas based on the plain conversational text. In this task, a best-matched persona is searched out from candidates given the conversational text. This is a many-to-many semantic matching task because both contexts and personas in SPD are composed of multiple sentences. The long-term dependency and the dynamic redundancy among these sentences increase the difficulty of this task. We build a dataset for SPD, dubbed as Persona Match on Persona-Chat (PMPC). Furthermore, we evaluate several baseline models and propose utterance-to-profile (U2P) matching networks for this task. The U2P models operate at a fine granularity which treat both contexts and personas as sets of multiple sequences. Then, each sequence pair is scored and an interpretable overall score is obtained for a context-persona pair through aggregation. Evaluation results show that the U2P models outperform their baseline counterparts significantly.

DuRecDial 2.0: A Bilingual Parallel Corpus for Conversational Recommendation

Zeming Liu et al.

In this paper, we provide a bilingual parallel human-to-human recommendation dialog dataset (DuRecDial 2.0) to enable researchers to explore a challenging task of multilingual and cross-lingual conversational recommendation. The difference between DuRecDial 2.0 and existing conversational recommendation datasets is that the data item (Profile, Goal, Knowledge, Context, Response) in DuRecDial 2.0 is annotated in two languages, both English and Chinese, while other datasets are built with the setting of a single language. We collect 8.2k dialogs aligned across English and Chinese languages (16.5k dialogs and 255k utterances in total) that are annotated by crowdsourced workers with strict quality control procedure. We then build monolingual, multilingual, and cross-lingual conversational recommendation baselines on DuRecDial 2.0. Experiment results show that the use of additional English data can bring performance improvement for Chinese conversational recommendation, indicating the benefits of DuRecDial 2.0. Finally, this dataset provides a challenging testbed for future studies of monolingual, multilingual, and cross-lingual conversational recommendation.

CR-Walker: Tree-Structured Graph Reasoning and Dialog Acts for Conversational Recommendation

Wenchang Ma, Ryuichi Takanobu, and Minlie Huang

Growing interests have been attracted in Conversational Recommender Systems (CRS), which explore user preference through conversational interactions in order to make appropriate recommendation. However, there

is still a lack of ability in existing CRS to (1) traverse multiple reasoning paths over background knowledge to introduce relevant items and attributes, and (2) arrange selected entities appropriately under current system intents to control response generation. To address these issues, we propose CR-Walker in this paper, a model that performs tree-structured reasoning on a knowledge graph, and generates informative dialog acts to guide language generation. The unique scheme of tree-structured reasoning views the traversed entity at each hop as part of dialog acts to facilitate language generation, which links how entities are selected and expressed. Automatic and human evaluations show that CR-Walker can arrive at more accurate recommendation, and generate more informative and engaging responses.

Dynamic Knowledge Distillation for Pre-trained Language Models
Lei Li et al.

Knowledge distillation~(KD) has been proved effective for compressing large-scale pre-trained language models. However, existing methods conduct KD statically, e.g., the student model aligns its output distribution to that of a selected teacher model on the pre-defined training dataset. In this paper, we explore whether a dynamic knowledge distillation that empowers the student to adjust the learning procedure according to its competency, regarding the student performance and learning efficiency. We explore the dynamical adjustments on three aspects: teacher model adoption, data selection, and KD objective adaptation. Experimental results show that (1) proper selection of teacher model can boost the performance of student model; (2) conducting KD with 10% informative instances achieves comparable performance while greatly accelerates the training; (3) the student performance can be boosted by adjusting the supervision contribution of different alignment objective. We find dynamic knowledge distillation is promising and provide discussions on potential future directions towards more efficient KD methods.

PermuteFormer: Efficient Relative Position Encoding for Long Sequences
Peng Chen

A recent variation of Transformer, Performer, scales Transformer to longer sequences with a linear attention mechanism. However, it is not compatible with relative position encoding, which has advantages over absolute position encoding. In this paper, we discuss possible ways to add relative position encoding to Performer. Based on the analysis, we propose PermuteFormer, a Performer-based model with relative position encoding that scales linearly on long sequences. PermuteFormer applies position-dependent transformation on queries and keys to encode positional information into the attention module. This transformation is carefully crafted so that the final output of self-attention is not affected by absolute positions of tokens. PermuteFormer introduces negligible computational overhead by design that it runs as fast as Performer. We evaluate PermuteFormer on Long-Range Arena, a dataset for long sequences, as well as WikiText-103, a language modeling dataset. The experiments show that PermuteFormer uniformly improves the performance of Performer with almost no computational overhead and outperforms vanilla Transformer on most of the tasks.

SOM-NCSCM : An Efficient Neural Chinese Sentence Compression Model Enhanced with Self-Organizing Map
Kangli Zi et al.

Sentence Compression (SC), which aims to shorten sentences while retaining important words that express the essential meanings, has been studied for many years in many languages, especially in English. However, improvements on Chinese SC task are still quite few due to several difficulties: scarce of parallel corpora, different segmentation granularity of Chinese sentences, and imperfect performance of syntactic analyses. Furthermore, entire neural Chinese SC models have been under-investigated so far. In this work, we construct an SC dataset of Chinese colloquial sentences from a real-life question answering system in the telecommunication domain, and then, we propose a neural Chinese SC model enhanced with a Self-Organizing Map (SOM-NCSCM), to gain a valuable insight from the data and improve the performance of the whole neural Chinese SC model in a valid manner. Experimental results show that our SOM-NCSCM can significantly benefit from the deep investigation of similarity among data, and achieve a promising F1 score of 89.655 and BLEU4 score of 70.116, which also provides a baseline for further research on Chinese SC task.

Distilling Linguistic Context for Language Model Compression
Geondo Park, Gyeongman Kim, and Eunho Yang

A computationally expensive and memory intensive neural network lies behind the recent success of language representation learning. Knowledge distillation, a major technique for deploying such a vast language model in resource-scarce environments, transfers the knowledge on individual word representations learned without restrictions. In this paper, inspired by the recent observations that language representations are relatively positioned and have more semantic knowledge as a whole, we present a new knowledge distillation objective for language representation learning that transfers the contextual knowledge via two types of relationships across representations: Word Relation and Layer Transforming Relation. Unlike other recent distillation techniques for the language models, our contextual distillation does not have any restrictions on architectural changes between teacher and student. We validate the effectiveness of our method on challenging benchmarks of language understanding tasks, not only in architectures of various sizes but also in combination with DynaBERT, the recently proposed adaptive size pruning method.

IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization
Fajri Koto, Jey Han Lau, and Timothy Baldwin

We present IndoBERTweet, the first large-scale pretrained model for Indonesian Twitter that is trained by extending a monolingually-trained Indonesian BERT model with additive domain-specific vocabulary. We focus in particular on efficient model adaptation under vocabulary mismatch, and benchmark different ways of initializing the BERT embedding layer for new word types. We find that initializing with the average BERT subword embedding makes pretraining five times faster, and is more effective than proposed methods for vocabulary adaptation in terms of extrinsic evaluation over seven Twitter-based datasets.

Universal-KD: Attention-based Output-Grounded Intermediate Layer Knowledge Distillation

Yimeng Wu et al.

Intermediate layer matching is shown as an effective approach for improving knowledge distillation (KD). However, this technique applies matching in the hidden spaces of two different networks (i.e. student and teacher), which lacks clear interpretability. Moreover, intermediate layer KD cannot easily deal with other problems such as layer mapping search and architecture mismatch (i.e. it requires the teacher and student to be of the same model type). To tackle the aforementioned problems all together, we propose Universal-KD to match intermediate layers of the teacher and the student in the output space (by adding pseudo classifiers on intermediate layers) via the attention-based layer projection. By doing this, our unified approach has three merits: (i) it can be flexibly combined with current intermediate layer distillation techniques to improve their results (ii) the pseudo classifiers of the teacher can be deployed instead of extra expensive teacher assistant networks to address the capacity gap problem in KD which is a common issue when the gap between the size of the teacher and student networks becomes too large; (iii) it can be used in cross-architecture intermediate layer KD. We did comprehensive experiments in distilling BERT-base into BERT-4, RoBERTa-large into DistilRoBERTa and BERT-base into CNN and LSTM-based models. Results on the GLUE tasks show that our approach is able to outperform other KD techniques.

Harms of Gender Exclusivity and Challenges in Non-Binary Representation in Language Technologies

Sunipa Dev et al.

Gender is widely discussed in the context of language tasks and when examining the stereotypes propagated by language models. However, current discussions primarily treat gender as binary, which can perpetuate harms such as the cyclical erasure of non-binary gender identities. These harms are driven by model and dataset biases, which are consequences of the non-recognition and lack of understanding of non-binary genders in society. In this paper, we explain the complexity of gender and language around it, and survey non-binary persons to understand harms associated with the treatment of gender as binary in English language technologies. We also detail how current language representations (e.g., GloVe, BERT) capture and perpetuate these harms and related challenges that need to be acknowledged and addressed for representations to equitably encode gender information.

Are Gender-Neutral Queries Really Gender-Neutral? Mitigating Gender Bias in Image Search

Jialu Wang, Yang Liu, and Xin Wang

Internet search affects people's cognition of the world, so mitigating biases in search results and learning fair models is imperative for social good. We study a unique gender bias in image search in this work: the search images are often gender-imbalanced for gender-neutral natural language queries. We diagnose two typical image search models, the specialized model trained on in-domain datasets and the generalized representation model pre-trained on massive image and text data across the internet. Both models suffer from severe gender bias. Therefore, we introduce two novel debiasing approaches: an in-processing fair sampling method to address the gender imbalance issue for training models, and a post-processing feature clipping method base on mutual information to debias multimodal representations of pre-trained models. Extensive experiments on MS-COCO and Flickr30K benchmarks show that our methods significantly reduce the gender bias in image search models.

Fairness-aware Class Imbalanced Learning

Shivashankar Subramanian et al.

Class imbalance is a common challenge in many NLP tasks, and has clear connections to bias, in that bias in training data often leads to higher accuracy for majority groups at the expense of minority groups. However there has traditionally been a disconnect between research on class-imbalanced learning and mitigating bias, and only recently have the two been looked at through a common lens. In this work we evaluate long-tail learning methods for tweet sentiment and occupation classification, and extend a margin-loss based approach with methods to enforce fairness. We empirically show through controlled experiments that the proposed approaches help mitigate both class imbalance and demographic biases.

Mitigating Language-Dependent Ethnic Bias in BERT

Jaimeen Ahn and Alice Oh

In this paper, we study ethnic bias and how it varies across languages by analyzing and mitigating ethnic bias in monolingual BERT for English, German, Spanish, Korean, Turkish, and Chinese. To observe and quantify ethnic bias, we develop a novel metric called Categorical Bias score. Then we propose two methods for mitigation; first using a multilingual model, and second using contextual word alignment of two monolingual

models. We compare our proposed methods with monolingual BERT and show that these methods effectively alleviate the ethnic bias. Which of the two methods works better depends on the amount of NLP resources available for that language. We additionally experiment with Arabic and Greek to verify that our proposed methods work for a wider variety of languages.

Mathematical Word Problem Generation from Commonsense Knowledge Graph and Equations
Tianqiao Liu et al.

There is an increasing interest in the use of mathematical word problem (MWP) generation in educational assessment. Different from standard natural question generation, MWP generation needs to maintain the underlying mathematical operations between quantities and variables, while at the same time ensuring the relevance between the output and the given topic. To address above problem, we develop an end-to-end neural model to generate diverse MWPs in real-world scenarios from commonsense knowledge graph and equations. The proposed model (1) learns both representations from edge-enhanced Levi graphs of symbolic equations and commonsense knowledge; (2) automatically fuses equation and commonsense knowledge information via a self-planning module when generating the MWPs. Experiments on an educational gold-standard set and a large-scale generated MWP set show that our approach is superior on the MWP generation task, and it outperforms the SOTA models in terms of both automatic evaluation metrics, i.e., BLEU-4, ROUGE-L, Self-BLEU, and human evaluation metrics, i.e., equation relevance, topic relevance, and language coherence. To encourage reproducible results, we make our code and MWP dataset public available at `https://github.com/tal-ai/MaKE_EMNLP2021`.

Revisiting Pivot-Based Paraphrase Generation: Language Is Not the Only Optional Pivot
Yitao Cai, Yue Cao, and Xiaojun Wan

Paraphrases refer to texts that convey the same meaning with different expression forms. Pivot-based methods, also known as the round-trip translation, have shown promising results in generating high-quality paraphrases. However, existing pivot-based methods all rely on language as the pivot, where large-scale, high-quality parallel bilingual texts are required. In this paper, we explore the feasibility of using semantic and syntactic representations as the pivot for paraphrase generation. Concretely, we transform a sentence into a variety of different semantic or syntactic representations (including AMR, UD, and latent semantic representation), and then decode the sentence back from the semantic representations. We further explore a pretraining-based approach to compress the pipeline process into an end-to-end framework. We conduct experiments comparing different approaches with different kinds of pivots. Experimental results show that taking AMR as pivot can obtain paraphrases with better quality than taking language as the pivot. The end-to-end framework can reduce semantic shift when language is used as the pivot. Besides, several unsupervised pivot-based methods can generate paraphrases with similar quality as the supervised sequence-to-sequence model, which indicates that parallel data of paraphrases may not be necessary for paraphrase generation.

DiscoDVT: Generating Long Text with Discourse-Aware Discrete Variational Transformer
Haozhe Ji and Minlie Huang

Despite the recent advances in applying pre-trained language models to generate high-quality texts, generating long passages that maintain long-range coherence is yet challenging for these models. In this paper, we propose DiscoDVT, a discourse-aware discrete variational Transformer to tackle the incoherence issue. DiscoDVT learns a discrete variable sequence that summarizes the global structure of the text and then applies it to guide the generation process at each decoding step. To further embed discourse-aware information into the discrete latent representations, we introduce an auxiliary objective to model the discourse relations within the text. We conduct extensive experiments on two open story generation datasets and demonstrate that the latent codes learn meaningful correspondence to the discourse structures that guide the model to generate long texts with better long-range coherence.

Open-domain clarification question generation without question examples
Julia White et al.

An overarching goal of natural language processing is to enable machines to communicate seamlessly with humans. However, natural language can be ambiguous or unclear. In cases of uncertainty, humans engage in an interactive process known as repair: asking questions and seeking clarification until their uncertainty is resolved. We propose a framework for building a visually grounded question-asking model capable of producing polar (yes-no) clarification questions to resolve misunderstandings in dialogue. Our model uses an expected information gain objective to derive informative questions from an off-the-shelf image captioner without requiring any supervised question-answer data. We demonstrate our model's ability to pose questions that improve communicative success in a goal-oriented 20 questions game with synthetic and human answerers.

Continual Learning for Grounded Instruction Generation by Observing Human Following Behavior
Noriyuki Kojima, Alane Suhr, and Yoav Artzi

We study continual learning for natural language instruction generation, by observing human users' instruction execution. We focus on a collaborative scenario, where the system both acts and delegates tasks to human users using natural language. We compare user execution of generated instructions to the original system intent as

an indication to the system's success communicating its intent. We show how to use this signal to improve the system's ability to generate instructions via contextual bandit learning. In interaction with real users, our system demonstrates dramatic improvements in its ability to generate language over time.

Logic-level Evidence Retrieval and Graph-based Verification Network for Table-based Fact Verification
Qi Shi et al.

Table-based fact verification task aims to verify whether the given statement is supported by the given semi-structured table. Symbolic reasoning with logical operations plays a crucial role in this task. Existing methods leverage programs that contain rich logical information to enhance the verification process. However, due to the lack of fully supervised signals in the program generation process, spurious programs can be derived and employed, which leads to the inability of the model to catch helpful logical operations. To address the aforementioned problems, in this work, we formulate the table-based fact verification task as an evidence retrieval and reasoning framework, proposing the Logic-level Evidence Retrieval and Graph-based Verification network (LERGV). Specifically, we first retrieve logic-level program-like evidence from the given table and statement as supplementary evidence for the table. After that, we construct a logic-level graph to capture the logical relations between entities and functions in the retrieved evidence, and design a graph-based verification network to perform logic-level graph-based reasoning based on the constructed graph to classify the final entailment relation. Experimental results on the large-scale benchmark TABFACT show the effectiveness of the proposed approach.

A Partition Filter Network for Joint Entity and Relation Extraction
Zhiheng Yan et al.

In joint entity and relation extraction, existing work either sequentially encode task-specific features, leading to an imbalance in inter-task feature interaction where features extracted later have no direct contact with those that come first. Or they encode entity features and relation features in a parallel manner, meaning that feature representation learning for each task is largely independent of each other except for input sharing. We propose a partition filter network to model two-way interaction between tasks properly, where feature encoding is decomposed into two steps: partition and filter. In our encoder, we leverage two gates: entity and relation gate, to segment neurons into two task partitions and one shared partition. The shared partition represents inter-task information valuable to both tasks and is evenly shared across two tasks to ensure proper two-way interaction. The task partitions represent intra-task information and are formed through concerted efforts of both gates, making sure that encoding of task-specific features is dependent upon each other. Experiment results on six public datasets show that our model performs significantly better than previous approaches. In addition, contrary to what previous work has claimed, our auxiliary experiments suggest that relation prediction is contributory to named entity prediction in a non-negligible way. The source code can be found at https://github.com/Coopercoppers/PFN.

TEBNER: Domain Specific Named Entity Recognition with Type Expanded Boundary-aware Network
Zheng Fang et al.

To alleviate label scarcity in Named Entity Recognition (NER) task, distantly supervised NER methods are widely applied to automatically label data and identify entities. Although the human effort is reduced, the generated incomplete and noisy annotations pose new challenges for learning effective neural models. In this paper, we propose a novel dictionary extension method which extracts new entities through the type expanded model. Moreover, we design a multi-granularity boundary-aware network which detects entity boundaries from both local and global perspectives. We conduct experiments on different types of datasets, the results show that our model outperforms previous state-of-the-art distantly supervised systems and even surpasses the supervised models.

Set Generation Networks for End-to-End Knowledge Base Population
Dianbo Sui et al.

The task of knowledge base population (KBP) aims to discover facts about entities from texts and expand a knowledge base with these facts. Previous studies shape end-to-end KBP as a machine translation task, which is required to convert unordered fact into a sequence according to a pre-specified order. However, the facts stated in a sentence are unordered in essence. In this paper, we formulate end-to-end KBP as a direct set generation problem, avoiding considering the order of multiple facts. To solve the set generation problem, we propose networks featured by transformers with non-autoregressive parallel decoding. Unlike previous approaches that use an autoregressive decoder to generate facts one by one, the proposed networks can directly output the final set of facts in one shot. Furthermore, to train the networks, we also design a set-based loss that forces unique predictions via bipartite matching. Compared with cross-entropy loss that highly penalizes small shifts in fact order, the proposed bipartite matching loss is invariant to any permutation of predictions. Benefiting from getting rid of the burden of predicting the order of multiple facts, our proposed networks achieve state-of-the-art (SoTA) performance on two benchmark datasets.

Knowing False Negatives: An Adversarial Training Method for Distantly Supervised Relation Extraction
Kailong Hao, Botao Yu, and Wei Hu

Distantly supervised relation extraction (RE) automatically aligns unstructured text with relation instances in a knowledge base (KB). Due to the incompleteness of current KBs, sentences implying certain relations may be annotated as N/A instances, which causes the so-called false negative (FN) problem. Current RE methods usually overlook this problem, inducing improper biases in both training and testing procedures. To address this issue, we propose a two-stage approach. First, it finds out possible FN samples by heuristically leveraging the memory mechanism of deep neural networks. Then, it aligns those unlabeled data with the training data into a unified feature space by adversarial training to assign pseudo labels and further utilize the information contained in them. Experiments on two wildly-used benchmark datasets demonstrate the effectiveness of our approach.

Fine-grained Entity Typing via Label Reasoning
Qing Liu et al.

Conventional entity typing approaches are based on independent classification paradigms, which make them difficult to recognize inter-dependent, long-tailed and fine-grained entity types. In this paper, we argue that the implicitly entailed extrinsic and intrinsic dependencies between labels can provide critical knowledge to tackle the above challenges. To this end, we propose Label Reasoning Network(LRN), which sequentially reasons fine-grained entity labels by discovering and exploiting label dependencies knowledge entailed in the data. Specifically, LRN utilizes an auto-regressive network to conduct deductive reasoning and a bipartite attribute graph to conduct inductive reasoning between labels, which can effectively model, learn and reason complex label dependencies in a sequence-to-set, end-to-end manner. Experiments show that LRN achieves the state-of-the-art performance on standard ultra fine-grained entity typing benchmarks, and can also resolve the long tail label problem effectively.

Structure-Augmented Keyphrase Generation
Jihyuk Kim et al.

This paper studies the keyphrase generation (KG) task for scenarios where structure plays an important role. For example, a scientific publication consists of a short title and a long body, where the title can be used for de-emphasizing unimportant details in the body. Similarly, for short social media posts (, tweets), scarce context can be augmented from titles, though often missing. Our contribution is generating/augmenting structure then injecting these information in the encoding, using existing keyphrases of other documents, complementing missing/incomplete titles. We propose novel structure-augmented document encoding approaches that consist of the following two phases: The first phase, generating structure, extends the given document with related but absent keyphrases, augmenting missing context. The second phase, encoding structure, builds a graph of keyphrases and the given document to obtain the structure-aware representation of the augmented text. Our empirical results validate that our proposed structure augmentation and augmentation-aware encoding/decoding can improve KG for both scenarios, outperforming the state-of-the-art.

Progressive Adversarial Learning for Bootstrapping: A Case Study on Entity Set Expansion
Lingyong Yan, Xianpei Han, and Le Sun

Bootstrapping has become the mainstream method for entity set expansion. Conventional bootstrapping methods mostly define the expansion boundary using seed-based distance metrics, which heavily depend on the quality of selected seeds and are hard to be adjusted due to the extremely sparse supervision. In this paper, we propose BootstrapGAN, a new learning method for bootstrapping which jointly models the bootstrapping process and the boundary learning process in a GAN framework. Specifically, the expansion boundaries of different bootstrapping iterations are learned via different discriminator networks; the bootstrapping network is the generator to generate new positive entities, and the discriminator networks identify the expansion boundaries by trying to distinguish the generated entities from known positive entities. By iteratively performing the above adversarial learning, the generator and the discriminators can reinforce each other and be progressively refined along the whole bootstrapping process. Experiments show that BootstrapGAN achieves the new state-of-the-art entity set expansion performance.

Maximal Clique Based Non-Autoregressive Open Information Extraction
Bowen Yu et al.

Open Information Extraction (OpenIE) aims to discover textual facts from a given sentence. In essence, the facts contained in plain text are unordered. However, the popular OpenIE systems usually output facts sequentially in the way of predicting the next fact conditioned on the previous decoded ones, which enforce an unnecessary order on the facts and involve the error accumulation between autoregressive steps. To break this bottleneck, we propose MacroIE, a novel non-autoregressive framework for OpenIE. MacroIE firstly constructs a fact graph based on the table filling scheme, in which each node denotes a fact element, and an edge links two nodes that belong to the same fact. Then OpenIE can be reformulated as a non-parametric process of finding maximal cliques from the graph. It directly outputs the final set of facts in one go, thus getting rid of the burden of predicting fact order, as well as the error propagation between facts. Experiments conducted on two benchmark datasets show that our proposed model significantly outperforms current state-of-the-art methods, beats the previous systems by as much as 5.7 absolute gain in F1 score.

A Relation-Oriented Clustering Method for Open Relation Extraction
Jun Zhao et al.

The clustering-based unsupervised relation discovery method has gradually become one of the important methods of open relation extraction (OpenRE). However, high-dimensional vectors can encode complex linguistic information which leads to the problem that the derived clusters cannot explicitly align with the relational semantic classes. In this work, we propose a relation-oriented clustering model and use it to identify the novel relations in the unlabeled data. Specifically, to enable the model to learn to cluster relational data, our method leverages the readily available labeled data of pre-defined relations to learn a relation-oriented representation. We minimize distance between the instance with same relation by gathering the instances towards their corresponding relation centroids to form a cluster structure, so that the learned representation is cluster-friendly. To reduce the clustering bias on predefined classes, we optimize the model by minimizing a joint objective on both labeled and unlabeled data. Experimental results show that our method reduces the error rate by 29.2% and 15.7%, on two datasets respectively, compared with current SOTA methods.

Learning Logic Rules for Document-Level Relation Extraction

Dongyu Ru et al.

Document-level relation extraction aims to identify relations between entities in a whole document. Prior efforts to capture long-range dependencies have relied heavily on implicitly powerful representations learned through (graph) neural networks, which makes the model less transparent. To tackle this challenge, in this paper, we propose LogiRE, a novel probabilistic model for document-level relation extraction by learning logic rules. LogiRE treats logic rules as latent variables and consists of two modules: a rule generator and a relation extractor. The rule generator is to generate logic rules potentially contributing to final predictions, and the relation extractor outputs final predictions based on the generated logic rules. Those two modules can be efficiently optimized with the expectation-maximization (EM) algorithm. By introducing logic rules into neural networks, LogiRE can explicitly capture long-range dependencies as well as enjoy better interpretation. Empirical results show that significantly outperforms several strong baselines in terms of relation performance and logical consistency. Our code is available at https://github.com/rudongyu/LogiRE.

Variational Deep Logic Network for Joint Inference of Entities and Relations

Wenya Wang and Sinno Jialin Pan

Currently, deep learning models have been widely adopted and achieved promising results on various application domains. Despite their intriguing performance, most deep learning models function as black boxes, lacking explicit reasoning capabilities and explanations, which are usually essential for complex problems. Take joint inference in information extraction as an example. This task requires the identification of multiple structured knowledge from texts, which is inter-correlated, including entities, events, and the relationships between them. Various deep neural networks have been proposed to jointly perform entity extraction and relation prediction, which only propagate information implicitly via representation learning. However, they fail to encode the intensive correlations between entity types and relations to enforce their coexistence. On the other hand, some approaches adopt rules to explicitly constrain certain relational facts, although the separation of rules with representation learning usually restrains the approaches with error propagation. Moreover, the predefined rules are inflexible and might result in negative effects when data is noisy. To address these limitations, we propose a variational deep logic network that incorporates both representation learning and relational reasoning via the variational EM algorithm. The model consists of a deep neural network to learn high-level features with implicit interactions via the self-attention mechanism and a relational logic network to explicitly exploit target interactions. These two components are trained interactively to bring the best of both worlds. We conduct extensive experiments ranging from fine-grained sentiment terms extraction, end-to-end relation prediction, to end-to-end event extraction to demonstrate the effectiveness of our proposed method.

Ultra-High Dimensional Sparse Representations with Binarization for Efficient Text Retrieval

Kyoung-Rok Jang et al.

The semantic matching capabilities of neural information retrieval can ameliorate synonymy and polysemy problems of symbolic approaches. However, neural models' dense representations are more suitable for re-ranking, due to their inefficiency. Sparse representations, either in symbolic or latent form, are more efficient with an inverted index. Taking the merits of the sparse and dense representations, we propose an ultra-high dimensional (UHD) representation scheme equipped with directly controllable sparsity. UHD's large capacity and minimal noise and interference among the dimensions allow for binarized representations, which are highly efficient for storage and search. Also proposed is a bucketing method, where the embeddings from multiple layers of BERT are selected/merged to represent diverse linguistic aspects. We test our models with MS MARCO and TREC CAR, showing that our models outperforms other sparse models.

Neural Attention-Aware Hierarchical Topic Model

YUAN JIN et al.

Neural topic models (NTMs) apply deep neural networks to topic modelling. Despite their success, NTMs generally ignore two important aspects: (1) only document-level word count information is utilized for the training, while more fine-grained sentence-level information is ignored, and (2) external semantic knowledge regarding documents, sentences and words are not exploited for the training. To address these issues, we propose a variational autoencoder (VAE) NTM model that jointly reconstructs the sentence and document word counts using combinations of bag-of-words (BoW) topical embeddings and pre-trained semantic embeddings. The pre-trained embeddings are first transformed into a common latent topical space to align their semantics with the BoW embeddings. Our model also features hierarchical KL divergence to leverage embeddings of

each document to regularize those of their sentences, paying more attention to semantically relevant sentences. Both quantitative and qualitative experiments have shown the efficacy of our model in 1) lowering the reconstruction errors at both the sentence and document levels, and 2) discovering more coherent topics from real-world datasets.

Incorporating Residual and Normalization Layers into Analysis of Masked Language Models

Goro Kobayashi et al.

Transformer architecture has become ubiquitous in the natural language processing field. To interpret the Transformer-based models, their attention patterns have been extensively analyzed. However, the Transformer architecture is not only composed of the multi-head attention; other components can also contribute to Transformers' progressive performance. In this study, we extended the scope of the analysis of Transformers from solely the attention patterns to the whole attention block, i.e., multi-head attention, residual connection, and layer normalization. Our analysis of Transformer-based masked language models shows that the token-to-token interaction performed via attention has less impact on the intermediate representations than previously assumed. These results provide new intuitive explanations of existing reports; for example, discarding the learned attention patterns tends not to adversely affect the performance. The codes of our experiments are publicly available.

Lifelong Explainer for Lifelong Learners

Xuelin Situ et al.

Lifelong Learning (LL) black-box models are dynamic in that they keep learning from new tasks and constantly update their parameters. Owing to the need to utilize information from previously seen tasks, and capture commonalities in potentially diverse data, it is hard for automatic explanation methods to explain the outcomes of these models. In addition, existing explanation methods, e.g., LIME, which are computationally expensive when explaining a static black-box model, are even more inefficient in the LL setting. In this paper, we propose a novel Lifelong Explanation (LLE) approach that continuously trains a student explainer under the supervision of a teacher – an arbitrary explanation algorithm – on different tasks undertaken in LL. We also leverage the Experience Replay (ER) mechanism to prevent catastrophic forgetting in the student explainer. Our experiments comparing LLE to three baselines on text classification tasks show that LLE can enhance the stability of the explanations for all seen tasks and maintain the same level of faithfulness to the black-box model as the teacher, while being up to 10^2 times faster at test time. Our ablation study shows that the ER mechanism in our LLE approach enhances the learning capabilities of the student explainer. Our code is available at https://github.com/situsnow/LLE.

On the Transferability of Adversarial Attacks against Neural Text Classifier

Liping Yuan et al.

Deep neural networks are vulnerable to adversarial attacks, where a small perturbation to an input alters the model prediction. In many cases, malicious inputs intentionally crafted for one model can fool another model. In this paper, we present the first study to systematically investigate the transferability of adversarial examples for text classification models and explore how various factors, including network architecture, tokenization scheme, word embedding, and model capacity, affect the transferability of adversarial examples. Based on these studies, we propose a genetic algorithm to find an ensemble of models that can be used to induce adversarial examples to fool almost all existing models. Such adversarial examples reflect the defects of the learning process and the data bias in the training set. Finally, we derive word replacement rules that can be used for model diagnostics from these adversarial examples.

SELFEXPLAIN: A Self-Explaining Architecture for Neural Text Classifiers

Dheeraj Rajagopal et al.

We introduce SelfExplain, a novel self-explaining model that explains a text classifier's predictions using phrase-based concepts. SelfExplain augments existing neural classifiers by adding (1) a globally interpretable layer that identifies the most influential concepts in the training set for a given sample and (2) a locally interpretable layer that quantifies the contribution of each local input concept by computing a relevance score relative to the predicted label. Experiments across five text-classification datasets show that SelfExplain facilitates interpretability without sacrificing performance. Most importantly, explanations from SelfExplain show sufficiency for model predictions and are perceived as adequate, trustworthy and understandable by human judges compared to existing widely-used baselines.

Predicting emergent linguistic compositions through time: Syntactic frame extension via multimodal chaining

Lei Yu and Yang Xu

Natural language relies on a finite lexicon to express an unbounded set of emerging ideas. One result of this tension is the formation of new compositions, such that existing linguistic units can be combined with emerging items into novel expressions. We develop a framework that exploits the cognitive mechanisms of chaining and multimodal knowledge to predict emergent compositional expressions through time. We present the syntactic frame extension model (SFEM) that draws on the theory of chaining and knowledge from "percept", "concept", and "language" to infer how verbs extend their frames to form new compositions with existing and novel nouns. We evaluate SFEM rigorously on the 1) modalities of knowledge and 2) categorization models of chaining,

in a syntactically parsed English corpus over the past 150 years. We show that multimodal SFEM predicts newly emerged verb syntax and arguments substantially better than competing models using purely linguistic or unimodal knowledge. We find support for an exemplar view of chaining as opposed to a prototype view and reveal how the joint approach of multimodal chaining may be fundamental to the creation of literal and figurative language uses including metaphor and metonymy.

Knowledge-Aware Meta-learning for Low-Resource Text Classification
Huaxiu Yao et al.

Meta-learning has achieved great success in leveraging the historical learned knowledge to facilitate the learning process of the new task. However, merely learning the knowledge from the historical tasks, adopted by current meta-learning algorithms, may not generalize well to testing tasks when they are not well-supported by training tasks. This paper studies a low-resource text classification problem and bridges the gap between meta-training and meta-testing tasks by leveraging the external knowledge bases. Specifically, we propose KGML to introduce additional representation for each sentence learned from the extracted sentence-specific knowledge graph. The extensive experiments on three datasets demonstrate the effectiveness of KGML under both supervised adaptation and unsupervised adaptation settings.

Raise a Child in Large Language Model: Towards Effective and Generalizable Fine-tuning
Runxin Xu et al.

Recent pretrained language models extend from millions to billions of parameters. Thus the need to fine-tune an extremely large pretrained model with a limited training corpus arises in various downstream tasks. In this paper, we propose a straightforward yet effective fine-tuning technique, Child-Tuning, which updates a subset of parameters (called child network) of large pretrained models via strategically masking out the gradients of the non-child network during the backward process. Experiments on various downstream tasks in GLUE benchmark show that Child-Tuning consistently outperforms the vanilla fine-tuning by 1.5~8.6 average score among four different pretrained models, and surpasses the prior fine-tuning techniques by 0.6~1.3 points. Furthermore, empirical results on domain transfer and task transfer show that Child-Tuning can obtain better generalization performance by large margins.

Efficient Contrastive Learning via Novel Data Augmentation and Curriculum Learning
Seonghyeon Ye, Jiseon Kim, and Alice Oh

We introduce EfficientCL, a memory-efficient continual pretraining method that applies contrastive learning with novel data augmentation and curriculum learning. For data augmentation, we stack two types of operation sequentially: cutoff and PCA jittering. While pretraining steps proceed, we apply curriculum learning by incrementing the augmentation degree for each difficulty step. After data augmentation is finished, contrastive learning is applied on projected embeddings of original and augmented examples. When finetuned on GLUE benchmark, our model outperforms baseline models, especially for sentence-level tasks. Additionally, this improvement is capable with only 70% of computational memory compared to the baseline model.

Neuralizing Regular Expressions for Slot Filling
Chengyue Jiang, Zijian Jin, and Kewei Tu

Neural models and symbolic rules such as regular expressions have their respective merits and weaknesses. In this paper, we study the integration of the two approaches for the slot filling task by converting regular expressions into neural networks. Specifically, we first convert regular expressions into a special form of finite-state transducers, then unfold its approximate inference algorithm as a bidirectional recurrent neural model that performs slot filling via sequence labeling. Experimental results show that our model has superior zero-shot and few-shot performance and stays competitive when there are sufficient training data.

Towards Zero-Shot Knowledge Distillation for Natural Language Processing
Ahmad Rashid et al.

Knowledge distillation (KD) is a common knowledge transfer algorithm used for model compression across a variety of deep learning based natural language processing (NLP) solutions. In its regular manifestations, KD requires access to the teacher's training data for knowledge transfer to the student network. However, privacy concerns, data regulations and proprietary reasons may prevent access to such data. We present, to the best of our knowledge, the first work on Zero-shot Knowledge Distillation for NLP, where the student learns from the much larger teacher without any task specific data. Our solution combines out-of-domain data and adversarial training to learn the teacher's output distribution. We investigate six tasks from the GLUE benchmark and demonstrate that we can achieve between 75% and 92% of the teacher's classification score (accuracy or F1) while compressing the model 30 times.

Provable Limitations of Acquiring Meaning from Ungrounded Form: What will Future Language Models Understand?
William Cooper Merrill et al.

Language models trained on billions of tokens have recently led to unprecedented results on many NLP tasks. This success raises the question of whether, in principle, a system can ever "understand" raw text without access to some form of grounding. We formally investigate the abilities of ungrounded systems to acquire meaning. Our analysis focuses on the role of "assertions": textual contexts that provide indirect clues about the underlying semantics. We study whether assertions enable a system to emulate representations preserving

semantic relations like equivalence. We find that assertions enable semantic emulation of languages that satisfy a strong notion of semantic transparency. However, for classes of languages where the same expression can take different values in different contexts, we show that emulation can become uncomputable. Finally, we discuss differences between our formal model and natural language, exploring how our results generalize to a modal setting and other semantic relations. Together, our results suggest that assertions in code or language do not provide sufficient signal to fully emulate semantic representations. We formalize ways in which ungrounded language models appear to be fundamentally limited in their ability to "understand".

Differentiable Subset Pruning of Transformer Heads
Jiaoda Li, Ryan Cotterell, and Mrinmaya Sachan

Multi-head attention, a collection of several attention mechanisms that independently attend to different parts of the input, is the key ingredient in the Transformer (Vaswani et al., 2017). Recent work has shown, however, that a large proportion of the heads in a Transformer's multi-head attention mechanism can be safely pruned away without significantly harming the performance of the model; such pruning leads to models that are noticeably smaller and faster in practice. Our work introduces a new head pruning technique that we term differentiable subset pruning. Intuitively, our method learns per-head importance variables and then enforces a user-specified hard constraint on the number of unpruned heads. The importance variables are learned via stochastic gradient descent. We conduct experiments on natural language inference and machine translation; we show that differentiable subset pruning significantly improves over the previously proposed head pruning techniques of Voita et al. (2019) and Michel et al. (2019).

Uncertainty-Aware Balancing for Multilingual and Multi-Domain Neural Machine Translation Training
Minghao Wu et al.

Learning multilingual and multi-domain translation model is challenging as the heterogeneous and imbalanced data make the model converge inconsistently over different corpora in real world. One common practice is to adjust the share of each corpus in the training, so that the learning process is balanced and low-resource cases can benefit from the high resource ones. However, automatic balancing methods usually depend on the intra- and inter-dataset characteristics, which is usually agnostic or requires human priors. In this work, we propose an approach, MultiUAT, that dynamically adjusts the training data usage based on the model's uncertainty on a small set of trusted clean data for multi-corpus machine translation. We experiments with two classes of uncertainty measures on multilingual (16 languages with 4 settings) and multi-domain settings (4 for in-domain and 2 for out-of-domain on English-German translation) and demonstrate our approach MultiUAT substantially outperforms its baselines, including both static and dynamic strategies. We analyze the cross-domain transfer and show the deficiency of static and similarity based methods.

ERNIE-M: Enhanced Multilingual Representation by Aligning Cross-lingual Semantics with Monolingual Corpora
Xuan Ouyang et al.

Recent studies have demonstrated that pre-trained cross-lingual models achieve impressive performance in downstream cross-lingual tasks. This improvement benefits from learning a large amount of monolingual and parallel corpora. Although it is generally acknowledged that parallel corpora are critical for improving the model performance, existing methods are often constrained by the size of parallel corpora, especially for low-resource languages. In this paper, we propose Ernie-M, a new training method that encourages the model to align the representation of multiple languages with monolingual corpora, to overcome the constraint that the parallel corpus size places on the model performance. Our key insight is to integrate back-translation into the pre-training process. We generate pseudo-parallel sentence pairs on a monolingual corpus to enable the learning of semantic alignments between different languages, thereby enhancing the semantic modeling of cross-lingual models. Experimental results show that Ernie-M outperforms existing cross-lingual models and delivers new state-of-the-art results in various cross-lingual downstream tasks. The codes and pre-trained models will be made publicly available.

mT6: Multilingual Pretrained Text-to-Text Transformer with Translation Pairs
Zewen Chi et al.

Multilingual T5 pretrains a sequence-to-sequence model on massive monolingual texts, which has shown promising results on many cross-lingual tasks. In this paper, we improve multilingual text-to-text transfer Transformer with translation pairs (mT6). Specifically, we explore three cross-lingual text-to-text pre-training tasks, namely, machine translation, translation pair span corruption, and translation span corruption. In addition, we propose a partially non-autoregressive objective for text-to-text pre-training. We evaluate the methods on seven multilingual benchmark datasets, including sentence classification, named entity recognition, question answering, and abstractive summarization. Experimental results show that the proposed mT6 improves cross-lingual transferability over mT5.

Cross Attention Augmented Transducer Networks for Simultaneous Translation
Dan Liu et al.

This paper proposes a novel architecture, Cross Attention Augmented Transducer (CAAT), for simultaneous translation. The framework aims to jointly optimize the policy and translation models. To effectively consider all possible READ-WRITE simultaneous translation action paths, we adapt the online automatic speech recog-

nition (ASR) model, RNN-T, but remove the strong monotonic constraint, which is critical for the translation task to consider reordering. To make CAAT work, we introduce a novel latency loss whose expectation can be optimized by a forward-backward algorithm. We implement CAAT with Transformer while the general CAAT architecture can also be implemented with other attention-based encoder-decoder frameworks. Experiments on both speech-to-text (S2T) and text-to-text (T2T) simultaneous translation tasks show that CAAT achieves significantly better latency-quality trade-offs compared to the state-of-the-art simultaneous translation approaches.

Translating Headers of Tabular Data: A Pilot Study of Schema Translation
Kunrui Zhu et al.

Schema translation is the task of automatically translating headers of tabular data from one language to another. High-quality schema translation plays an important role in cross-lingual table searching, understanding and analysis. Despite its importance, schema translation is not well studied in the community, and state-of-the-art neural machine translation models cannot work well on this task because of two intrinsic differences between plain text and tabular data: morphological difference and context difference. To facilitate the research study, we construct the first parallel dataset for schema translation, which consists of 3,158 tables with 11,979 headers written in 6 different languages, including English, Chinese, French, German, Spanish, and Japanese. Also, we propose the first schema translation model called CAST, which is a header-to-header neural machine translation model augmented with schema context. Specifically, we model a target header and its context as a directed graph to represent their entity types and relations. Then CAST encodes the graph with a relational-aware transformer and uses another transformer to decode the header in the target language. Experiments on our dataset demonstrate that CAST significantly outperforms state-of-the-art neural machine translation models. Our dataset will be released at https://github.com/microsoft/ContextualSP.

Towards Making the Most of Dialogue Characteristics for Neural Chat Translation
Yunlong Liang et al.

Neural Chat Translation (NCT) aims to translate conversational text between speakers of different languages. Despite the promising performance of sentence-level and context-aware neural machine translation models, there still remain limitations in current NCT models because the inherent dialogue characteristics of chat, such as dialogue coherence and speaker personality, are neglected. In this paper, we propose to promote the chat translation by introducing the modeling of dialogue characteristics into the NCT model. To this end, we design four auxiliary tasks including monolingual response generation, cross-lingual response generation, next utterance discrimination, and speaker identification. Together with the main chat translation task, we optimize the enhanced NCT model through the training objectives of all these tasks. By this means, the NCT model can be enhanced by capturing the inherent dialogue characteristics, thus generating more coherent and speaker-relevant translations. Comprehensive experiments on four language directions (English<->German and English<->Chinese) verify the effectiveness and superiority of the proposed approach.

It Is Not As Good As You Think! Evaluating Simultaneous Machine Translation on Interpretation Data
Jinming Zhao et al.

Most existing simultaneous machine translation (SiMT) systems are trained and evaluated on offline translation corpora. We argue that SiMT systems should be trained and tested on real interpretation data. To illustrate this argument, we propose an interpretation test set and conduct a realistic evaluation of SiMT trained on offline translations. Our results, on our test set along with 3 existing smaller scale language pairs, highlight the difference of up-to 13.83 BLEU score when SiMT models are evaluated on translation vs interpretation data. In the absence of interpretation training data, we propose a translation-to-interpretation (T2I) style transfer method which allows converting existing offline translations into interpretation-style data, leading to up-to 2.8 BLEU improvement. However, the evaluation gap remains notable, calling for constructing large-scale interpretation corpora better suited for evaluating and developing SiMT systems.

Translation-based Supervision for Policy Generation in Simultaneous Neural Machine Translation
Ashkan Alinejad, Hassan S. Shavarani, and Anoop Sarkar

In simultaneous machine translation, finding an agent with the optimal action sequence of reads and writes that maintain a high level of translation quality while minimizing the average lag in producing target tokens remains an extremely challenging problem. We propose a novel supervised learning approach for training an agent that can detect the minimum number of reads required for generating each target token by comparing simultaneous translations against full-sentence translations during training to generate oracle action sequences. These oracle sequences can then be used to train a supervised model for action generation at inference time. Our approach provides an alternative to current heuristic methods in simultaneous translation by introducing a new training objective, which is easier to train than previous attempts at training the agent using reinforcement learning techniques for this task. Our experimental results show that our novel training method for action generation produces much higher quality translations while minimizing the average lag in simultaneous translation.

Learning Kernel-Smoothed Machine Translation with Retrieved Examples
Qingnan Jiang et al.

How to effectively adapt neural machine translation (NMT) models according to emerging cases without re-

training? Despite the great success of neural machine translation, updating the deployed models online remains a challenge. Existing non-parametric approaches that retrieve similar examples from a database to guide the translation process are promising but are prone to overfit the retrieved examples. However, non-parametric methods are prone to overfit the retrieved examples. In this work, we propose to learn Kernel-Smoothed Translation with Example Retrieval (KSTER), an effective approach to adapt neural machine translation models online. Experiments on domain adaptation and multi-domain machine translation datasets show that even without expensive retraining, KSTER is able to achieve improvement of 1.1 to 1.5 BLEU scores over the best existing online adaptation methods. The code and trained models are released at https://github.com/jiangqn/KSTER.

AligNART: Non-autoregressive Neural Machine Translation by Jointly Learning to Estimate Alignment and Translate
Jongyoon Song, Sungwon Kim, and Sungroh Yoon

Non-autoregressive neural machine translation (NART) models suffer from the multi-modality problem which causes translation inconsistency such as token repetition. Most recent approaches have attempted to solve this problem by implicitly modeling dependencies between outputs. In this paper, we introduce AligNART, which leverages full alignment information to explicitly reduce the modality of the target distribution. AligNART divides the machine translation task into (i) alignment estimation and (ii) translation with aligned decoder inputs, guiding the decoder to focus on simplified one-to-one translation. To alleviate the alignment estimation problem, we further propose a novel alignment decomposition method. Our experiments show that AligNART outperforms previous non-iterative NART models that focus on explicit modality reduction on WMT14 EnDe and WMT16 Ro→En. Furthermore, AligNART achieves BLEU scores comparable to those of the state-of-the-art connectionist temporal classification based models on WMT14 EnDe. We also observe that AligNART effectively addresses the token repetition problem even without sequence-level knowledge distillation.

Meta Distant Transfer Learning for Pre-trained Language Models
Chengyu Wang et al.

With the wide availability of Pre-trained Language Models (PLMs), multi-task fine-tuning across domains has been extensively applied. For tasks related to distant domains with different class label sets, PLMs may memorize non-transferable knowledge for the target domain and suffer from negative transfer. Inspired by meta-learning, we propose the Meta Distant Transfer Learning (Meta-DTL) framework to learn the cross-task knowledge for PLM-based methods. Meta-DTL first employs task representation learning to mine implicit relations among multiple tasks and classes. Based on the results, it trains a PLM-based meta-learner to capture the transferable knowledge across tasks. The weighted maximum entropy regularizers are proposed to make meta-learner more task-agnostic and unbiased. Finally, the meta-learner can be fine-tuned to fit each task with better parameter initialization. We evaluate Meta-DTL using both BERT and ALBERT on seven public datasets. Experiment results confirm the superiority of Meta-DTL as it consistently outperforms strong baselines. We find that Meta-DTL is highly effective when very few data is available for the target task.

UniKER: A Unified Framework for Combining Embedding and Definite Horn Rule Reasoning for Knowledge Graph Inference
Ming Zhang and Yizhou Sun

Knowledge graph inference has been studied extensively due to its wide applications. It has been addressed by two lines of research, i.e., the more traditional logical rule reasoning and the more recent knowledge graph embedding (KGE). Several attempts have been made to combine KGE and logical rules for better knowledge graph inference. Unfortunately, they either simply treat logical rules as additional constraints into KGE loss or use probabilistic model to approximate the exact logical inference (i.e., MAX-SAT). Even worse, both approaches need to sample ground rules to tackle the scalability issue, as the total number of ground rules is intractable in practice, making them less effective in handling logical rules. In this paper, we propose a novel framework UniKER to address these challenges by restricting logical rules to be definite Horn rules, which can fully exploit the knowledge in logical rules and enable the mutual enhancement of logical rule-based reasoning and KGE in an extremely efficient way. Extensive experiments have demonstrated that our approach is superior to existing state-of-the-art algorithms in terms of both efficiency and effectiveness.

Voice Query Auto Completion
Raphael Tang et al.

Query auto completion (QAC) is the task of predicting a search engine user's final query from their intermediate, incomplete query. In this paper, we extend QAC to the streaming voice search setting, where automatic speech recognition systems produce intermediate transcriptions as users speak. Naively applying existing methods fails because the intermediate transcriptions often don't form prefixes or even substrings of the final transcription. To address this issue, we propose to condition QAC approaches on intermediate transcriptions to complete voice queries. We evaluate our models on a speech-enabled smart television with real-life voice search traffic, finding that this ASR-aware conditioning improves the completion quality. Our best method obtains an 18% relative improvement in mean reciprocal rank over previous methods.

AVocaDo: Strategy for Adapting Vocabulary to Downstream Domain
Jimin Hong et al.

During the fine-tuning phase of transfer learning, the pretrained vocabulary remains unchanged, while model parameters are updated. The vocabulary generated based on the pretrained data is suboptimal for downstream

data when domain discrepancy exists. We propose to consider the vocabulary as an optimizable parameter, allowing us to update the vocabulary by expanding it with domain specific vocabulary based on a tokenization statistic. Furthermore, we preserve the embeddings of the added words from overfitting to downstream data by utilizing knowledge learned from a pretrained language model with a regularization term. Our method achieved consistent performance improvements on diverse domains (i.e., biomedical, computer science, news, and reviews).

Wasserstein Selective Transfer Learning for Cross-domain Text Mining
Lingyun Feng et al.

Transfer learning (TL) seeks to improve the learning of a data-scarce target domain by using information from source domains. However, the source and target domains usually have different data distributions, which may lead to negative transfer. To alleviate this issue, we propose a Wasserstein Selective Transfer Learning (WSTL) method. Specifically, the proposed method considers a reinforced selector to select helpful data for transfer learning. We further use a Wasserstein-based discriminator to maximize the empirical distance between the selected source data and target data. The TL module is then trained to minimize the estimated Wasserstein distance in an adversarial manner and provides domain invariant features for the reinforced selector. We adopt an evaluation metric based on the performance of the TL module as delayed reward and a Wasserstein-based metric as immediate rewards to guide the reinforced selector learning. Compared with the competing TL approaches, the proposed method selects data samples that are closer to the target domain. It also provides better state features and reward signals that lead to better performance with faster convergence. Extensive experiments on three real-world text mining tasks demonstrate the effectiveness of the proposed method.

Jointly Learning to Repair Code and Generate Commit Message
Jiaqi Bai et al.

We propose a novel task of jointly repairing program codes and generating commit messages. Code repair and commit message generation are two essential and related tasks for software development. However, existing work usually performs the two tasks independently. We construct a multilingual triple dataset including buggy code, fixed code, and commit messages for this novel task. We first introduce a cascaded method with two models, one is to generate the fixed code first, and the other generates the commit message based on the fixed and original codes. We enhance the cascaded method with different training approaches, including the teacher-student method, the multi-task method, and the back-translation method. To deal with the error propagation problem of the cascaded method, we also propose a joint model that can both repair the program code and generate the commit message in a unified framework. Massive experiments on our constructed buggy-fixed-commit dataset reflect the challenge of this task and that the enhanced cascaded model and the proposed joint model significantly outperform baselines in both quality of code and commit messages.

Can Language Models be Biomedical Knowledge Bases?
Mujeen Sung et al.

Pre-trained language models (LMs) have become ubiquitous in solving various natural language processing (NLP) tasks. There has been increasing interest in what knowledge these LMs contain and how we can extract that knowledge, treating LMs as knowledge bases (KBs). While there has been much work on probing LMs in the general domain, there has been little attention to whether these powerful LMs can be used as domain-specific KBs. To this end, we create the BioLAMA benchmark, which is comprised of 49K biomedical factual knowledge triples for probing biomedical LMs. We find that biomedical LMs with recently proposed probing methods can achieve up to 18.51% Acc5 on retrieving biomedical knowledge. Although this seems promising given the task difficulty, our detailed analyses reveal that most predictions are highly correlated with prompt templates without any subjects, hence producing similar results on each relation and hindering their capabilities to be used as domain-specific KBs. We hope that BioLAMA can serve as a challenging benchmark for biomedical factual probing.

Abstract, Rationale, Stance: A Joint Model for Scientific Claim Verification
Zhiwei Zhang et al.

Scientific claim verification can help the researchers to easily find the target scientific papers with the sentence evidence from a large corpus for the given claim. Some existing works propose pipeline models on the three tasks of abstract retrieval, rationale selection and stance prediction. Such works have the problems of error propagation among the modules in the pipeline and lack of sharing valuable information among modules. We thus propose an approach, named as ARSJoint, that jointly learns the modules for the three tasks with a machine reading comprehension framework by including claim information. In addition, we enhance the information exchanges and constraints among tasks by proposing a regularization term between the sentence attention scores of abstract retrieval and the estimated outputs of rational selection. The experimental results on the benchmark dataset SciFact show that our approach outperforms the existing works.

Exploring Methods for Generating Feedback Comments for Writing Learning
Kazuaki Hanawa, Ryo Nagata, and Kentaro Inui

The task of generating explanatory notes for language learners is known as feedback comment generation. Although various generation techniques are available, little is known about which methods are appropriate for this task. Nagata (2019) demonstrates the effectiveness of neural-retrieval-based methods in generating feedback comments for preposition use. Retrieval-based methods have limitations in that they can only output

feedback comments existing in a given training data. Furthermore, feedback comments can be made on other grammatical and writing items than preposition use, which is still unaddressed. To shed light on these points, we investigate a wider range of methods for generating many feedback comments in this study. Our close analysis of the type of task leads us to investigate three different architectures for comment generation: (i) a neural-retrieval-based method as a baseline, (ii) a pointer-generator-based generation method as a neural seq2seq method, (iii) a retrieve-and-edit method, a hybrid of (i) and (ii). Intuitively, the pointer-generator should outperform neural-retrieval, and retrieve-and-edit should perform best. However, in our experiments, this expectation is completely overturned. We closely analyze the results to reveal the major causes of these counter-intuitive results and report on our findings from the experiments.

Lexically-Aware Semi-Supervised Learning for OCR Post-Correction
Shruti Rijhwani et al.

Much of the existing linguistic data in many languages of the world is locked away in non-digitized books and documents. Optical character recognition (OCR) can be used to produce digitized text, and previous work has demonstrated the utility of neural post-correction methods that improve the results of general-purpose OCR systems on recognition of less-well-resourced languages. However, these methods rely on manually curated post-correction data, which are relatively scarce compared to the non-annotated raw images that need to be digitized. In this paper, we present a semi-supervised learning method that makes it possible to utilize these raw images to improve performance, specifically through the use of self-training, a technique where a model is iteratively trained on its own outputs. In addition, to enforce consistency in the recognized vocabulary, we introduce a lexically-aware decoding method that augments the neural post-correction model with a count-based language model constructed from the recognized texts, implemented using weighted finite-state automata (WFSA) for efficient and effective decoding. Results on four endangered languages demonstrate the utility of the proposed method, with relative error reductions of 15-29%, where we find the combination of self-training and lexically-aware decoding essential for achieving consistent improvements.

Local Word Discovery for Interactive Transcription
William Lane and Steven Bird

Human expertise and the participation of speech communities are essential factors in the success of technologies for low-resource languages. Accordingly, we propose a new computational task which is tuned to the available knowledge and interests in an Indigenous community, and which supports the construction of high quality texts and lexicons. The task is illustrated for Kunwinjku, a morphologically-complex Australian language. We combine a finite state implementation of a published grammar with a partial lexicon, and apply this to a noisy phone representation of the signal. We locate known lexemes in the signal and use the morphological transducer to build these out into hypothetical, morphologically-complex words for human validation. We show that applying a single iteration of this method results in a relative transcription density gain of 17%. Further, we find that 75% of breath groups in the test set receive at least one correct partial or full-word suggestion.

Segment, Mask, and Predict: Augmenting Chinese Word Segmentation with Self-Supervision
Mieradilijiang Maimaiti et al.

Recent state-of-the-art (SOTA) effective neural network methods and fine-tuning methods based on pre-trained models (PTM) have been used in Chinese word segmentation (CWS), and they achieve great results. However, previous works focus on training the models with the fixed corpus at every iteration. The intermediate generated information is also valuable. Besides, the robustness of the previous neural methods is limited by the large-scale annotated data. There are a few noises in the annotated corpus. Limited efforts have been made by previous studies to deal with such problems. In this work, we propose a self-supervised CWS approach with a straightforward and effective architecture. First, we train a word segmentation model and use it to generate the segmentation results. Then, we use a revised masked language model (MLM) to evaluate the quality of the segmentation results based on the predictions of the MLM. Finally, we leverage the evaluations to aid the training of the segmenter by improved minimum risk training. Experimental results show that our approach outperforms previous methods on 9 different CWS datasets with single criterion training and multiple criteria training and achieves better robustness.

TransferNet: An Effective and Transparent Framework for Multi-hop Question Answering over Relation Graph
Jiaxin Shi et al.

Multi-hop Question Answering (QA) is a challenging task because it requires precise reasoning with entity relations at every step towards the answer. The relations can be represented in terms of labels in knowledge graph (e.g., spouse) or text in text corpus (e.g., they have been married for 26 years). Existing models usually infer the answer by predicting the sequential relation path or aggregating the hidden graph features. The former is hard to optimize, and the latter lacks interpretability. In this paper, we propose TransferNet, an effective and transparent model for multi-hop QA, which supports both label and text relations in a unified framework. TransferNet jumps across entities at multiple steps. At each step, it attends to different parts of the question, computes activated scores for relations, and then transfer the previous entity scores along activated relations in a differentiable way. We carry out extensive experiments on three datasets and demonstrate that TransferNet surpasses the state-of-the-art models by a large margin. In particular, on MetaQA, it achieves 100% accuracy in 2-hop and 3-hop questions. By qualitative analysis, we show that TransferNet has transparent and interpretable

intermediate results.

Generative Context Pair Selection for Multi-hop Question Answering
Dheeru Dua et al.

Compositional reasoning tasks such as multi-hop question answering require models to learn how to make latent decisions using only weak supervision from the final answer. Crowdsourced datasets gathered for these tasks, however, often contain only a slice of the underlying task distribution, which can induce unanticipated biases such as shallow word overlap between the question and context. Recent works have shown that discriminative training results in models that exploit these underlying biases to achieve a better held-out performance, without learning the right way to reason. We propose a generative context selection model for multi-hop QA that reasons about how the given question could have been generated given a context pair and not just independent contexts. We show that on HotpotQA, while being comparable to the state-of-the-art answering performance, our proposed generative passage selection model has a better performance (4.9% higher than baseline) on adversarial held-out set which tests robustness of model's multi-hop reasoning capabilities.

Have You Seen That Number? Investigating Extrapolation in Question Answering Models
Jeonghwan Kim et al.

Numerical reasoning in machine reading comprehension (MRC) has shown drastic improvements over the past few years. While the previous models for numerical MRC are able to interpolate the learned numerical reasoning capabilities, it is not clear whether they can perform just as well on numbers unseen in the training dataset. Our work rigorously tests state-of-the-art models on DROP, a numerical MRC dataset, to see if they can handle passages that contain out-of-range numbers. One of the key findings is that the models fail to extrapolate to unseen numbers. Presenting numbers as digit-by-digit input to the model, we also propose the *E-digit* number form that alleviates the lack of extrapolation in models and reveals the need to treat numbers differently from regular words in the text. Our work provides a valuable insight into the numerical MRC models and the way to represent number forms in MRC.

WebSRC: A Dataset for Web-Based Structural Reading Comprehension
Xingyu Chen et al.

Web search is an essential way for humans to obtain information, but it's still a great challenge for machines to understand the contents of web pages. In this paper, we introduce the task of web-based structural reading comprehension. Given a web page and a question about it, the task is to find an answer from the web page. This task requires a system not only to understand the semantics of texts but also the structure of the web page. Moreover, we proposed WebSRC, a novel Web-based Structural Reading Comprehension dataset. WebSRC consists of 400K question-answer pairs, which are collected from 6.4K web pages with corresponding HTML source code, screenshots, and metadata. Each question in WebSRC requires a certain structural understanding of a web page to answer, and the answer is either a text span on the web page or yes/no. We evaluate various strong baselines on our dataset to show the difficulty of our task. We also investigate the usefulness of structural information and visual features. Our dataset and baselines have been publicly available.

Back-Training excels Self-Training at Unsupervised Domain Adaptation of Question Generation and Passage Retrieval
Devang Kulshreshtha et al.

In this work, we introduce back-training, an alternative to self-training for unsupervised domain adaptation (UDA). While self-training generates synthetic training data where natural inputs are aligned with noisy outputs, back-training results in natural outputs aligned with noisy inputs. This significantly reduces the gap between target domain and synthetic data distribution, and reduces model overfitting to source domain. We run UDA experiments on question generation and passage retrieval from the Natural Questions domain to machine learning and biomedical domains. We find that back-training vastly outperforms self-training by a mean improvement of 7.8 BLEU-4 points on generation, and 17.6% top-20 retrieval accuracy across both domains. We further propose consistency filters to remove low-quality synthetic data before training. We also release a new domain-adaptation dataset - MLQuestions containing 35K unaligned questions, 50K unaligned passages, and 3K aligned question-passage pairs.

FinQA: A Dataset of Numerical Reasoning over Financial Data
Zhiyu Chen et al.

The sheer volume of financial statements makes it difficult for humans to access and analyze a business's financials. Robust numerical reasoning likewise faces unique challenges in this domain. In this work, we focus on answering deep questions over financial data, aiming to automate the analysis of a large corpus of financial documents. In contrast to existing tasks on general domain, the finance domain includes complex numerical reasoning and understanding of heterogeneous representations. To facilitate analytical progress, we propose a new large-scale dataset, FinQA, with Question-Answering pairs over Financial reports, written by financial experts. We also annotate the gold reasoning programs to ensure full explainability. We further introduce baselines and conduct comprehensive experiments in our dataset. The results demonstrate that popular, large, pre-trained models fall far short of expert humans in acquiring finance knowledge and in complex multi-step numerical reasoning on that knowledge. Our dataset – the first of its kind – should therefore enable significant, new community research into complex application domains. The dataset and code are publicly available athttps://github.com/czyssrs/FinQA.

FiD-Ex: Improving Sequence-to-Sequence Models for Extractive Rationale Generation
Kushal Lakhotia et al.

Natural language (NL) explanations of model predictions are gaining popularity as a means to understand and verify decisions made by large black-box pre-trained models, for tasks such as Question Answering (QA) and Fact Verification. Recently, pre-trained sequence to sequence (seq2seq) models have proven to be very effective in jointly making predictions, as well as generating NL explanations. However, these models have many shortcomings; they can fabricate explanations even for incorrect predictions, they are difficult to adapt to long input documents, and their training requires a large amount of labeled data. In this paper, we develop FiD-Ex, which addresses these shortcomings for seq2seq models by: 1) introducing sentence markers to eliminate explanation fabrication by encouraging extractive generation, 2) using the fusion-in-decoder architecture to handle long input contexts, and 3) intermediate fine-tuning on re-structured open domain QA datasets to improve few-shot performance. FiD-Ex significantly improves over prior work in terms of explanation metrics and task accuracy on five tasks from the ERASER explainability benchmark in both fully supervised and few-shot settings.

How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering
Zhengbao Jiang et al.

Recent works have shown that language models (LM) capture different types of knowledge regarding facts or common sense. However, because no model is perfect, they still fail to provide appropriate answers in many cases. In this paper, we ask the question "how can we know when language models know, with confidence, the answer to a particular query?" We examine this question from the point of view of calibration, the property of a probabilistic model's predicted probabilities actually being well correlated with the probabilities of correctness. We examine three strong generative models – T5, BART, and GPT-2 – and study whether their probabilities on QA tasks are well calibrated, finding the answer is a relatively emphatic no. We then examine methods to calibrate such models to make their confidence scores correlate better with the likelihood of correctness through fine-tuning, post-hoc probability modification, or adjustment of the predicted outputs or inputs. Experiments on a diverse range of datasets demonstrate the effectiveness of our methods. We also perform analysis to study the strengths and limitations of these methods, shedding light on further improvements that may be made in methods for calibrating LMs. We have released the code at https://github.com/jzbjyb/lm-calibration.

MKQA: A Linguistically Diverse Benchmark for Multilingual Open Domain Question Answering
Shayne Longpre, Yi Lu, and Joachim Daiber

Progress in cross-lingual modeling depends on challenging, realistic, and diverse evaluation sets. We introduce Multilingual Knowledge Questions and Answers (MKQA), an open-domain question answering evaluation set comprising 10k question-answer pairs aligned across 26 typologically diverse languages (260k question-answer pairs in total). The goal of this dataset is to provide a challenging benchmark for question answering quality across a wide set of languages. Answers are based on a language-independent data representation, making results comparable across languages and independent of language-specific passages. With 26 languages, this dataset supplies the widest range of languages to-date for evaluating question answering. We benchmark a wide variety of state-of-the-art methods in retrieval with generative and extractive question answering baselines, trained on Natural Questions, in zero shot and translation settings. Results indicate this dataset is challenging in all languages, and especially in low-resource languages.

CrossFit: A Few-shot Learning Challenge for Cross-task Generalization in NLP
Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren

Humans can learn a new language task efficiently with only few examples, by leveraging their knowledge obtained when learning prior tasks. In this paper, we explore whether and how such cross-task generalization ability can be acquired, and further applied to build better few-shot learners across diverse NLP tasks. We introduce CrossFit, a problem setup for studying cross-task generalization ability, which standardizes seen/unseen task partitions, data access during different learning stages, and the evaluation protocols. To instantiate different seen/unseen task partitions in CrossFit and facilitate in-depth analysis, we present the NLP Few-shot Gym, a repository of 160 diverse few-shot NLP tasks created from open-access NLP datasets and converted to a unified text-to-text format. Our analysis reveals that the few-shot learning ability on unseen tasks can be improved via an upstream learning stage using a set of seen tasks. We also observe that the selection of upstream learning tasks can significantly influence few-shot performance on unseen tasks, asking further analysis on task similarity and transferability.

IndoNLI: A Natural Language Inference Dataset for Indonesian
Rahmad Mahendra et al.

We present IndoNLI, the first human-elicited NLI dataset for Indonesian. We adapt the data collection protocol for MNLI and collect ~18K sentence pairs annotated by crowd workers and experts. The expert-annotated data is used exclusively as a test set. It is designed to provide a challenging test-bed for Indonesian NLI by explicitly incorporating various linguistic phenomena such as numerical reasoning, structural changes, idioms, or temporal and spatial reasoning. Experiment results show that XLM-R outperforms other pre-trained models in our data. The best performance on the expert-annotated data is still far below human performance (13.4% accuracy gap), suggesting that this test set is especially challenging. Furthermore, our analysis shows that our expert-annotated data is more diverse and contains fewer annotation artifacts than the crowd-annotated data.

We hope this dataset can help accelerate progress in Indonesian NLP research.

CodRED: A Cross-Document Relation Extraction Dataset for Acquiring Knowledge in the Wild

Yuan Yao et al.

Existing relation extraction (RE) methods typically focus on extracting relational facts between entity pairs within single sentences or documents. However, a large quantity of relational facts in knowledge bases can only be inferred across documents in practice. In this work, we present the problem of cross-document RE, making an initial step towards knowledge acquisition in the wild. To facilitate the research, we construct the first human-annotated cross-document RE dataset CodRED. Compared to existing RE datasets, CodRED presents two key challenges: Given two entities, (1) it requires finding the relevant documents that can provide clues for identifying their relations; (2) it requires reasoning over multiple documents to extract the relational facts. We conduct comprehensive experiments to show that CodRED is challenging to existing RE methods including strong BERT-based models.

AfroMT: Pretraining Strategies and Reproducible Benchmarks for Translation of 8 African Languages

Machel Reid et al.

Reproducible benchmarks are crucial in driving progress of machine translation research. However, existing machine translation benchmarks have been mostly limited to high-resource or well-represented languages. Despite an increasing interest in low-resource machine translation, there are no standardized reproducible benchmarks for many African languages, many of which are used by millions of speakers but have less digitized textual data. To tackle these challenges, we propose AfroMT, a standardized, clean, and reproducible machine translation benchmark for eight widely spoken African languages. We also develop a suite of analysis tools for system diagnosis taking into account the unique properties of these languages. Furthermore, we explore the newly considered case of low-resource focused pretraining and develop two novel data augmentation-based strategies, leveraging word-level alignment information and pseudo-monolingual data for pretraining multilingual sequence-to-sequence models. We demonstrate significant improvements when pretraining on 11 languages, with gains of up to 2 BLEU points over strong baselines. We also show gains of up to 12 BLEU points over cross-lingual transfer baselines in data-constrained scenarios. All code and pretrained models will be released as further steps towards larger reproducible benchmarks for African languages.

ParsiNLU: A Suite of Language Understanding Challenges for Persian

Daniel Khashabi et al.

Despite the progress made in recent years in addressing natural language understanding (NLU) challenges, the majority of this progress remains to be concentrated on resource-rich languages like English. This work focuses on Persian language, one of the widely spoken languages in the world, and yet there are few NLU datasets available for this language. The availability of high-quality evaluation datasets is a necessity for reliable assessment of the progress on different NLU tasks and domains. We introduce ParsiNLU, the first benchmark in Persian language that includes a range of language understanding tasks — Reading Comprehension, Textual Entailment, etc. These datasets are collected in a multitude of ways, often involving manual annotations by native speakers. This results in over 14.5k new instances across 6 distinct NLU tasks. Besides, we present the first results on state-of-the-art monolingual and multi-lingual pre-trained language models on this benchmark and compare them with human performance, which provides valuable insights into our ability to tackle natural language understanding challenges in Persian. We hope ParsiNLU fosters further research and advances in Persian language understanding.

Iterative Paraphrastic Augmentation with Discriminative Span Alignment

Benjamin Van Durme et al.

We introduce a novel paraphrastic augmentation strategy based on sentence-level lexically constrained paraphrasing and discriminative span alignment. Our approach allows for the large-scale expansion of existing datasets or the rapid creation of new datasets using a small, manually-produced seed corpus. We demonstrate our approach with experiments on the Berkeley FrameNet Project, a large-scale language understanding effort spanning more than two decades of human labor. With four days of training data collection for a span alignment model and one day of parallel compute, we automatically generate and release to the community 495,300 unique (Frame, Trigger) pairs in diverse sentential contexts, a roughly 50-fold expansion atop FrameNet v1.7. The resulting dataset is intrinsically and extrinsically evaluated in detail, showing positive results on a downstream task.

LM-Critic: Language Models for Unsupervised Grammatical Error Correction

Michihiro Yasunaga, Jure Leskovec, and Percy Liang

Grammatical error correction (GEC) requires a set of labeled ungrammatical / grammatical sentence pairs for training, but obtaining such annotation can be prohibitively expensive. Recently, the Break-It-Fix-It (BIFI) framework has demonstrated strong results on learning to repair a broken program without any labeled examples, but this relies on a perfect critic (e.g., a compiler) that returns whether an example is valid or not, which does not exist for the GEC task. In this work, we show how to leverage a pretrained language model (LM) in defining an LM-Critic, which judges a sentence to be grammatical if the LM assigns it a higher probability than its local perturbations. We apply this LM-Critic and BIFI along with a large set of unlabeled sentences

to bootstrap realistic ungrammatical / grammatical pairs for training a corrector. We evaluate our approach on GEC datasets on multiple domains (CoNLL-2014, BEA-2019, GMEG-wiki and GMEG-yahoo) and show that it outperforms existing methods in both the unsupervised setting (+7.7 F0.5) and the supervised setting (+0.5 F0.5).

Integrating Deep Event-Level and Script-Level Information for Script Event Prediction
Long Bai et al.

Scripts are structured sequences of events together with the participants, which are extracted from the texts. Script event prediction aims to predict the subsequent event given the historical events in the script. Two kinds of information facilitate this task, namely, the event-level information and the script-level information. At the event level, existing studies view an event as a verb with its participants, while neglecting other useful properties, such as the state of the participants. At the script level, most existing studies only consider a single event sequence corresponding to one common protagonist. In this paper, we propose a Transformer-based model, called MCPredictor, which integrates deep event-level and script-level information for script event prediction. At the event level, MCPredictor utilizes the rich information in the text to obtain more comprehensive event semantic representations. At the script-level, it considers multiple event sequences corresponding to different participants of the subsequent event. The experimental results on the widely-used New York Times corpus demonstrate the effectiveness and superiority of the proposed model.

PICARD: Parsing Incrementally for Constrained Auto-Regressive Decoding from Language Models
Torsten Scholak, Nathan Schucher, and Dzmitry Bahdanau

Large pre-trained language models for textual data have an unconstrained output space; at each decoding step, they can produce any of 10,000s of sub-word tokens. When fine-tuned to target constrained formal languages like SQL, these models often generate invalid code, rendering it unusable. We propose PICARD (code available at https://github.com/ElementAI/picard), a method for constraining auto-regressive decoders of language models through incremental parsing. PICARD helps to find valid output sequences by rejecting inadmissible tokens at each decoding step. On the challenging Spider and CoSQL text-to-SQL translation tasks, we show that PICARD transforms fine-tuned T5 models with passable performance into state-of-the-art solutions.

Semantics-Preserved Data Augmentation for Aspect-Based Sentiment Analysis
Ting-Wei Hsu et al.

Both the issues of data deficiencies and semantic consistency are important for data augmentation. Most of previous methods address the first issue, but ignore the second one. In the cases of aspect-based sentiment analysis, violation of the above issues may change the aspect and sentiment polarity. In this paper, we propose a semantics-preservation data augmentation approach by considering the importance of each word in a textual sequence according to the related aspects and sentiments. We then substitute the unimportant tokens with two replacement strategies without altering the aspect-level polarity. Our approach is evaluated on several publicly available sentiment analysis datasets and the real-world stock price/risk movement prediction scenarios. Experimental results show that our methodology achieves better performances in all datasets.

Joint Multi-modal Aspect-Sentiment Analysis with Auxiliary Cross-modal Relation Detection
Xincheng Ju et al.

Aspect terms extraction (ATE) and aspect sentiment classification (ASC) are two fundamental and fine-grained sub-tasks in aspect-level sentiment analysis (ALSA). In the textual analysis, joint extracting both aspect terms and sentiment polarities has been drawn much attention due to the better applications than individual sub-task. However, in the multi-modal scenario, the existing studies are limited to handle each sub-task independently, which fails to model the innate connection between the above two objectives and ignores the better applications. Therefore, in this paper, we are the first to jointly perform multi-modal ATE (MATE) and multi-modal ASC (MASC), and we propose a multi-modal joint learning approach with auxiliary cross-modal relation detection for multi-modal aspect-level sentiment analysis (MALSA). Specifically, we first build an auxiliary text-image relation detection module to control the proper exploitation of visual information. Second, we adopt the hierarchical framework to bridge the multi-modal connection between MATE and MASC, as well as separately visual guiding for each sub module. Finally, we can obtain all aspect-level sentiment polarities dependent on the jointly extracted specific aspects. Extensive experiments show the effectiveness of our approach against the joint textual approaches, pipeline and collapsed multi-modal approaches.

Solving Aspect Category Sentiment Analysis as a Text Generation Task
Jian Liu et al.

Aspect category sentiment analysis has attracted increasing research attention. The dominant methods make use of pre-trained language models by learning effective aspect category-specific representations, and adding specific output layers to its pre-trained representation. We consider a more direct way of making use of pre-trained language models, by casting the ACSA tasks into natural language generation tasks, using natural language sentences to represent the output. Our method allows more direct use of pre-trained knowledge in seq2seq language models by directly following the task setting during pre-training. Experiments on several benchmarks show that our method gives the best reported results, having large advantages in few-shot and

zero-shot settings.

Dimensional Emotion Detection from Categorical Emotion
Sungjoon Park et al.

We present a model to predict fine-grained emotions along the continuous dimensions of valence, arousal, and dominance (VAD) with a corpus with categorical emotion annotations. Our model is trained by minimizing the EMD (Earth Mover's Distance) loss between the predicted VAD score distribution and the categorical emotion distributions sorted along VAD, and it can simultaneously classify the emotion categories and predict the VAD scores for a given sentence. We use pre-trained RoBERTa-Large and fine-tune on three different corpora with categorical labels and evaluate on EmoBank corpus with VAD scores. We show that our approach reaches comparable performance to that of the state-of-the-art classifiers in categorical emotion classification and shows significant positive correlations with the ground truth VAD scores. Also, further training with supervision of VAD labels leads to improved performance especially when dataset is small. We also present examples of predictions of appropriate emotion words that are not part of the original annotations.

Learning Implicit Sentiment in Aspect-based Sentiment Analysis with Supervised Contrastive Pre-Training
Zhengyan Li et al.

Aspect-based sentiment analysis aims to identify the sentiment polarity of a specific aspect in product reviews. We notice that about 30% of reviews do not contain obvious opinion words, but still convey clear human-aware sentiment orientation, which is known as implicit sentiment. However, recent neural network-based approaches paid little attention to implicit sentiment entailed in the reviews. To overcome this issue, we adopt Supervised Contrastive Pre-training on large-scale sentiment-annotated corpora retrieved from in-domain language resources. By aligning the representation of implicit sentiment expressions to those with the same sentiment label, the pre-training process leads to better capture of both implicit and explicit sentiment orientation towards aspects in reviews. Experimental results show that our method achieves state-of-the-art performance on SemEval2014 benchmarks, and comprehensive analysis validates its effectiveness on learning implicit sentiment.

Progressive Self-Training with Discriminator for Aspect Term Extraction
Qianlong Wang et al.

Aspect term extraction aims to extract aspect terms from a review sentence that users have expressed opinions on. One of the remaining challenges for aspect term extraction resides in the lack of sufficient annotated data. While self-training is potentially an effective method to address this issue, the pseudo-labels it yields on unlabeled data could induce noise. In this paper, we use two means to alleviate the noise in the pseudo-labels. One is that inspired by the curriculum learning, we refine the conventional self-training to progressive self-training. Specifically, the base model infers pseudo-labels on a progressive subset at each iteration, where samples in the subset become harder and more numerous as the iteration proceeds. The other is that we use a discriminator to filter the noisy pseudo-labels. Experimental results on four SemEval datasets show that our model significantly outperforms the previous baselines and achieves state-of-the-art performance.

Chinese Opinion Role Labeling with Corpus Translation: A Pivot Study
Ranran Zhen et al.

Opinion Role Labeling (ORL), aiming to identify the key roles of opinion, has received increasing interest. Unlike most of the previous works focusing on the English language, in this paper, we present the first work of Chinese ORL. We construct a Chinese dataset by manually translating and projecting annotations from a standard English MPQA dataset. Then, we investigate the effectiveness of cross-lingual transfer methods, including model transfer and corpus translation. We exploit multilingual BERT with Contextual Parameter Generator and Adapter methods to examine the potentials of unsupervised cross-lingual learning and our experiments and analyses for both bilingual and multilingual transfers establish a foundation for the future research of this task.

Implicit Sentiment Analysis with Event-centered Text Representation
Deyu Zhou et al.

Implicit sentiment analysis, aiming at detecting the sentiment of a sentence without sentiment words, has become an attractive research topic in recent years. In this paper, we focus on event-centric implicit sentiment analysis that utilizes the sentiment-aware event contained in a sentence to infer its sentiment polarity. Most existing methods in implicit sentiment analysis simply view noun phrases or entities in text as events or indirectly model events with sophisticated models. Since events often trigger sentiments in sentences, we argue that this task would benefit from explicit modeling of events and event representation learning. To this end, we represent an event as the combination of its event type and the event triplet <subject, predicate, object>. Based on such event representation, we further propose a novel model with hierarchical tensor-based composition mechanism to detect sentiment in text. In addition, we present a dataset for event-centric implicit sentiment analysis where each sentence is labeled with the event representation described above. Experimental results on our constructed dataset and an existing benchmark dataset show the effectiveness of the proposed approach.

Reinforced Counterfactual Data Augmentation for Dual Sentiment Classification
Hao Chen, Rui Xia, and Jianfei Yu

Data augmentation and adversarial perturbation approaches have recently achieved promising results in solving the over-fitting problem in many natural language processing (NLP) tasks including sentiment classification. However, existing studies aimed to improve the generalization ability by augmenting the training data with synonymous examples or adding random noises to word embeddings, which cannot address the spurious association problem. In this work, we propose an end-to-end reinforcement learning framework, which jointly performs counterfactual data generation and dual sentiment classification. Our approach has three characteristics:1) the generator automatically generates massive and diverse antonymous sentences; 2) the discriminator contains a original-side sentiment predictor and an antonymous-side sentiment predictor, which jointly evaluate the quality of the generated sample and help the generator iteratively generate higher-quality antonymous samples; 3) the discriminator is directly used as the final sentiment classifier without the need to build an extra one. Extensive experiments show that our approach outperforms strong data augmentation baselines on several benchmark sentiment classification datasets. Further analysis confirms our approach's advantages in generating more diverse training samples and solving the spurious association problem in sentiment classification.

Interactive Machine Comprehension with Dynamic Knowledge Graphs

Xingdi Yuan

Interactive machine reading comprehension (iMRC) is machine comprehension tasks where knowledge sources are partially observable. An agent must interact with an environment sequentially to gather necessary knowledge in order to answer a question. We hypothesize that graph representations are good inductive biases, which can serve as an agent's memory mechanism in iMRC tasks. We explore four different categories of graphs that can capture text information at various levels. We describe methods that dynamically build and update these graphs during information gathering, as well as neural models to encode graph representations in RL agents. Extensive experiments on iSQuAD suggest that graph representations can result in significant performance improvements for RL agents.

LayoutReader: Pre-training of Text and Layout for Reading Order Detection

Zilong Wang et al.

Reading order detection is the cornerstone to understanding visually-rich documents (e.g., receipts and forms). Unfortunately, no existing work took advantage of advanced deep learning models because it is too laborious to annotate a large enough dataset. We observe that the reading order of WORD documents is embedded in their XML metadata; meanwhile, it is easy to convert WORD documents to PDFs or images. Therefore, in an automated manner, we construct ReadingBank, a benchmark dataset that contains reading order, text, and layout information for 500,000 document images covering a wide spectrum of document types. This first-ever large-scale dataset unleashes the power of deep neural networks for reading order detection. Specifically, our proposed LayoutReader captures the text and layout information for reading order prediction using the seq2seq model. It performs almost perfectly in reading order detection and significantly improves both open-source and commercial OCR engines in ordering text lines in their results in our experiments. The dataset and models are publicly available at https://aka.ms/layoutreader.

On Pursuit of Designing Multi-modal Transformer for Video Grounding

Meng Cao et al.

Video grounding aims to localize the temporal segment corresponding to a sentence query from an untrimmed video. Almost all existing video grounding methods fall into two frameworks: 1) Top-down model: It predefines a set of segment candidates and then conducts segment classification and regression. 2) Bottom-up model: It directly predicts frame-wise probabilities of the referential segment boundaries. However, all these methods are not end-to-end, i.e., they always rely on some time-consuming post-processing steps to refine predictions. To this end, we reformulate video grounding as a set prediction task and propose a novel end-to-end multi-modal Transformer model, dubbed as GTR. Specifically, GTR has two encoders for video and language encoding, and a cross-modal decoder for grounding prediction. To facilitate the end-to-end training, we use a Cubic Embedding layer to transform the raw videos into a set of visual tokens. To better fuse these two modalities in the decoder, we design a new Multi-head Cross-Modal Attention. The whole GTR is optimized via a Many-to-One matching loss. Furthermore, we conduct comprehensive studies to investigate different model design choices. Extensive results on three benchmarks have validated the superiority of GTR. All three typical GTR variants achieve record-breaking performance on all datasets and metrics, with several times faster inference speed.

Residual Adapters for Parameter-Efficient ASR Adaptation to Atypical and Accented Speech

Katrin Tomanek et al.

Automatic Speech Recognition (ASR) systems are often optimized to work best for speakers with canonical speech patterns. Unfortunately, these systems perform poorly when tested on atypical speech and heavily accented speech. It has previously been shown that personalization through model fine-tuning substantially improves performance. However, maintaining such large models per speaker is costly and difficult to scale. We show that by adding a relatively small number of extra parameters to the encoder layers via so-called residual adapter, we can achieve similar adaptation gains compared to model fine-tuning, while only updating a tiny fraction (less than 0.5%) of the model parameters. We demonstrate this on two speech adaptation tasks (atypical and accented speech) and for two state-of-the-art ASR architectures.

Broaden the Vision: Geo-Diverse Visual Commonsense Reasoning
Da Yin et al.

Commonsense is defined as the knowledge on which everyone agrees. However, certain types of commonsense knowledge are correlated with culture and geographic locations and they are only shared locally. For example, the scenes of wedding ceremonies vary across regions due to different customs influenced by historical and religious factors. Such regional characteristics, however, are generally omitted in prior work. In this paper, we construct a Geo-Diverse Visual Commonsense Reasoning dataset (GD-VCR) to test vision-and-language models' ability to understand cultural and geo-location-specific commonsense. In particular, we study two state-of-the-art Vision-and-Language models, VisualBERT and ViLBERT trained on VCR, a standard benchmark with images primarily from Western regions. We then evaluate how well the trained models can generalize to answering the questions in GD-VCR. We find that the performance of both models for non-Western regions including East Asia, South Asia, and Africa is significantly lower than that for Western region. We analyze the reasons behind the performance disparity and find that the performance gap is larger on QA pairs that: 1) are concerned with culture-related scenarios, e.g., weddings, religious activities, and festivals; 2) require high-level geo-diverse commonsense reasoning rather than low-order perception and recognition. Dataset and code are released at https://github.com/WadeYin9712/GD-VCR.

Region under Discussion for visual dialog
Mauricio Mazuecos et al.

Visual Dialog is assumed to require the dialog history to generate correct responses during a dialog. However, it is not clear from previous work how dialog history is needed for visual dialog. In this paper we define what it means for a visual question to require dialog history and we release a subset of the Guesswhat?! questions for which their dialog history completely changes their responses. We propose a novel interpretable representation that visually grounds dialog history: the Region under Discussion. It constrains the image's spatial features according to a semantic representation of the history inspired by the information structure notion of Question under Discussion.We evaluate the architecture on task-specific multimodal models and the visual transformer model LXMERT.

Reference-Centric Models for Grounded Collaborative Dialogue
Daniel Fried, Justin Chiu, and Dan Klein

We present a grounded neural dialogue model that successfully collaborates with people in a partially-observable reference game. We focus on a setting where two agents each observe an overlapping part of a world context and need to identify and agree on some object they share. Therefore, the agents should pool their information and communicate pragmatically to solve the task. Our dialogue agent accurately grounds referents from the partner's utterances using a structured reference resolver, conditions on these referents using a recurrent memory, and uses a pragmatic generation procedure to ensure the partner can resolve the references the agent produces. We evaluate on the OneCommon spatial grounding dialogue task (Udagawa and Aizawa 2019), involving a number of dots arranged on a board with continuously varying positions, sizes, and shades. Our agent substantially outperforms the previous state of the art for the task, obtaining a 20% relative improvement in successful task completion in self-play evaluations and a 50% relative improvement in success in human evaluations.

Effect of Visual Extensions on Natural Language Understanding in Vision-and-Language Models
Taichi Iki and Akiko Aizawa

A method for creating a vision-and-language (V&L) model is to extend a language model through structural modifications and V&L pre-training. Such an extension aims to make a V&L model inherit the capability of natural language understanding (NLU) from the original language model. To see how well this is achieved, we propose to evaluate V&L models using an NLU benchmark (GLUE). We compare five V&L models, including single-stream and dual-stream models, trained with the same pre-training. Dual-stream models, with their higher modality independence achieved by approximately doubling the number of parameters, are expected to preserve the NLU capability better. Our main finding is that the dual-stream scores are not much different than the single-stream scores, contrary to expectation. Further analysis shows that pre-training causes the performance drop in NLU tasks with few exceptions. These results suggest that adopting a single-stream structure and devising the pre-training could be an effective method for improving the maintenance of language knowledge in V&L extensions.

Generative Spoken Language Modeling from Raw Audio
Kushal Lakhotia et al.

We introduce Generative Spoken Language Modeling, the task of jointly learning the acoustic and linguistic characteristics of a language from raw audio (without text), and a set of metrics to automatically evaluate the learned representations at acoustic and linguistic levels for both encoding and generation. We set up baseline systems consisting of a discrete speech encoder (returning pseudo-text units), a generative language model (trained on pseudo-text), and a speech decoder (generating a waveform from pseudo-text) and validate the proposed metrics with human evaluation. Across three unsupervised speech encoders (CPC, wav2vec 2.0, HuBERT), we find that the number of discrete units (50, 100, or 200) matters in a task-dependent and encoderdependant way, and that some combinations approach text-based topline systems.

Low-Resource Dialogue Summarization with Domain-Agnostic Multi-Source Pretraining
Yicheng Zou et al.

With the rapid increase in the volume of dialogue data from daily life, there is a growing demand for dialogue summarization. Unfortunately, training a large summarization model is generally infeasible due to the inadequacy of dialogue data with annotated summaries. Most existing works for low-resource dialogue summarization directly pretrain models in other domains, e.g., the news domain, but they generally neglect the huge difference between dialogues and conventional articles. To bridge the gap between out-of-domain pretraining and in-domain fine-tuning, in this work, we propose a multi-source pretraining paradigm to better leverage the external summary data. Specifically, we exploit large-scale in-domain non-summary data to separately pretrain the dialogue encoder and the summary decoder. The combined encoder-decoder model is then pretrained on the out-of-domain summary data using adversarial critics, aiming to facilitate domain-agnostic summarization. The experimental results on two public datasets show that with only limited training data, our approach achieves competitive performance and generalizes well in different dialogue scenarios.

Fine-grained Factual Consistency Assessment for Abstractive Summarization Models
Sen Zhang, Jianwei Niu, and Chuyuan Wei

Factual inconsistencies existed in the output of abstractive summarization models with original documents are frequently presented. Fact consistency assessment requires the reasoning capability to find subtle clues to identify whether a model-generated summary is consistent with the original document. This paper proposes a fine-grained two-stage Fact Consistency assessment framework for Summarization models (SumFC). Given a document and a summary sentence, in the first stage, SumFC selects the top-K most relevant sentences with the summary sentence from the document. In the second stage, the model performs fine-grained consistency reasoning at the sentence level, and then aggregates all sentences' consistency scores to obtain the final assessment result. We get the training data pairs by data synthesis and adopt contrastive loss of data pairs to help the model identify subtle cues. Experiment results show that SumFC has made a significant improvement over the previous state-of-the-art methods. Our experiments also indicate that SumFC distinguishes detailed differences better.

A Statistical Analysis of Summarization Evaluation Metrics Using Resampling Methods
Daniel Deutsch, Rotem Dror, and Dan Roth

The quality of a summarization evaluation metric is quantified by calculating the correlation between its scores and human annotations across a large number of summaries. Currently, it is unclear how precise these correlation estimates are, nor whether differences between two metrics' correlations reflect a true difference or if it is due to mere chance. In this work, we address these two problems by proposing methods for calculating confidence intervals and running hypothesis tests for correlations using two resampling methods, bootstrapping and permutation. After evaluating which of the proposed methods is most appropriate for summarization through two simulation experiments, we analyze the results of applying these methods to several different automatic evaluation metrics across three sets of human annotations. We find that the confidence intervals are rather wide, demonstrating high uncertainty in the reliability of automatic metrics. Further, although many metrics fail to show statistical improvements over ROUGE, two recent works, QAEval and BERTScore, do in some evaluation settings.

Unsupervised Abstractive Opinion Summarization by Generating Sentences with Tree-Structured Topic Guidance
Masaru Isonuma et al.

This paper presents a novel unsupervised abstractive summarization method for opinionated texts. While the basic variational autoencoder-based models assume a unimodal Gaussian prior for the latent code of sentences, we alternate it with a recursive Gaussian mixture, where each mixture component corresponds to the latent code of a topic sentence and is mixed by a tree-structured topic distribution. By decoding each Gaussian component, we generate sentences with tree-structured topic guidance, where the root sentence conveys generic content, and the leaf sentences describe specific topics. Experimental results demonstrate that the generated topic sentences are appropriate as a summary of opinionated texts, which are more informative and cover more input contents than those generated by the recent unsupervised summarization model (Brazinskas et al., 2020). Furthermore, we demonstrate that the variance of latent Gaussians represents the granularity of sentences, analogous to Gaussian word embedding (Vilnis and McCallum, 2015).

Efficient Computation of Expectations under Spanning Tree Distributions
Ran Zmigrod, Tim Vieira, and Ryan Cotterell

We give a general framework for inference in spanning tree models. We propose unified algorithms for the important cases of first-order expectations and second-order expectations in edge-factored, non-projective spanning-tree models. Our algorithms exploit a fundamental connection between gradients and expectations, which allows us to derive efficient algorithms. These algorithms are easy to implement with or without automatic differentiation software. We motivate the development of our framework with several cautionary tales of previous research, which has developed numerous inefficient algorithms for computing expectations and their gradients. We demonstrate how our framework efficiently computes several quantities with known algorithms, including the expected attachment score, entropy, and generalized expectation criteria. As a bonus, we give algorithms for quantities that are missing in the literature, including the KL divergence. In all cases, our approach matches the efficiency of existing algorithms and, in several cases, reduces the runtime complexity by a factor of the sentence length. We validate the implementation of our framework through runtime experiments. We find our algorithms are up to 15 and 9 times faster than previous algorithms for computing the Shannon entropy and the gradient of the generalized expectation objective, respectively.

Joint Universal Syntactic and Semantic Parsing
Elias Stengel-Eskin et al.

While numerous attempts have been made to jointly parse syntax and semantics, high performance in one domain typically comes at the price of performance in the other. This trade-off contradicts the large body of research focusing on the rich interactions at the syntax-semantics interface. We explore multiple model architectures which allow us to exploit the rich syntactic and semantic annotations contained in the Universal Decompositional Semantics (UDS) dataset, jointly parsing Universal Dependencies and UDS to obtain state-of-the-art results in both formalisms. We analyze the behaviour of a joint model of syntax and semantics, finding patterns supported by linguistic theory at the syntax-semantics interface. We then investigate to what degree joint modeling generalizes to a multilingual setting, where we find similar trends across 8 languages.

A New Representation for Span-based CCG Parsing
Yoshihide Kato and Shigeki Matsubara

This paper proposes a new representation for CCG derivations. CCG derivations are represented as trees whose nodes are labeled with categories strictly restricted by CCG rule schemata. This characteristic is not suitable for span-based parsing models because they predict node labels independently. In other words, span-based models may generate invalid CCG derivations that violate the rule schemata. Our proposed representation decomposes CCG derivations into several independent pieces and prevents the span-based parsing models from violating the schemata. Our experimental result shows that an off-the-shelf span-based parser with our representation is comparable with previous CCG parsers.

Improving Sequence-to-Sequence Pre-training via Sequence Span Rewriting
Wangchunshu Zhou et al.

In this paper, we propose **S**equence **S**pan **R**ewriting (SSR), a self-supervised task for sequence-to-sequence (Seq2Seq) pre-training. SSR learns to refine the machine-generated imperfect text spans into ground truth text. SSR provides more fine-grained and informative supervision in addition to the original text-infilling objective. Compared to the prevalent text infilling objectives for Seq2Seq pre-training, SSR is naturally more consistent with many downstream generation tasks that require sentence rewriting (e.g., text summarization, question generation, grammatical error correction, and paraphrase generation). We conduct extensive experiments by using SSR to improve the typical Seq2Seq pre-trained model T5 in a continual pre-training setting and show substantial improvements over T5 on various natural language generation tasks.

Enhanced Language Representation with Label Knowledge for Span Extraction
Pan Yang et al.

Span extraction, aiming to extract text spans (such as words or phrases) from plain text, is a fundamental process in Information Extraction. Recent works introduce the label knowledge to enhance the text representation by formalizing the span extraction task into a question answering problem (QA Formalization), which achieves state-of-the-art performance. However, such a QA Formalization does not fully exploit the label knowledge and causes a dramatic decrease in efficiency of training/inference. To address those problems, we introduce a fresh paradigm to integrate label knowledge and further propose a novel model to explicitly and efficiently integrate label knowledge into text representations. Specifically, it encodes texts and label annotations independently and then integrates label knowledge into text representation with an elaborate-designed semantics fusion module. We conduct extensive experiments on three typical span extraction tasks: flat NER, nested NER, and event detection. The empirical results show that 1) our model achieves a new state-of-the-art performance on four benchmarks, and 2) reduces training time and inference time by 76% and 77% on average, respectively, compared with the QA Formalization paradigm.

Virtual Demo Session

N-LTP: An Open-source Neural Language Technology Platform for Chinese
Wanxiang Che et al.

We introduce N-LTP, an open-source neural language technology platform supporting six fundamental Chinese NLP tasks: lexical analysis (Chinese word segmentation, part-of-speech tagging, and named entity recognition), syntactic parsing (dependency parsing), and semantic parsing (semantic dependency parsing and semantic role labeling). Unlike the existing state-of-the-art toolkits, such as Stanza, that adopt an independent model for each task, N-LTP adopts the multi-task framework by using a shared pre-trained model, which has the advantage of capturing the shared knowledge across relevant Chinese tasks. In addition, a knowledge distillation method (Clark et al., 2019) where the single-task model teaches the multi-task model is further introduced to encourage the multi-task model to surpass its single-task teacher. Finally, we provide a collection of easy-to-use APIs and a visualization tool to make users to use and view the processing results more easily and directly. To the best of our knowledge, this is the first toolkit to support six Chinese NLP fundamental tasks. Source code, documentation, and pre-trained models are available at https://github.com/HIT-SCIR/ltp.

Athena 2.0: Contextualized Dialogue Management for an Alexa Prize SocialBot
Marilyn Walker et al.

Athena 2.0 is an Alexa Prize SocialBot that has been a finalist in the last two Alexa Prize Grand Challenges. One reason for Athena's success is its novel dialogue management strategy, which allows it to dynamically construct dialogues and responses from component modules, leading to novel conversations with every in-

teraction. Here we describe Athena's system design and performance in the Alexa Prize during the 20/21 competition. A live demo of Athena as well as video recordings will provoke discussion on the state of the art in conversational AI.

Open-Domain Question-Answering for COVID-19 and Other Emergent Domains
Sharon Levy et al.

Since late 2019, COVID-19 has quickly emerged as the newest biomedical domain, resulting in a surge of new information. As with other emergent domains, the discussion surrounding the topic has been rapidly changing, leading to the spread of misinformation. This has created the need for a public space for users to ask questions and receive credible, scientific answers. To fulfill this need, we turn to the task of open-domain question-answering, which we can use to efficiently find answers to free-text questions from a large set of documents. In this work, we present such a system for the emergent domain of COVID-19. Despite the small data size available, we are able to successfully train the system to retrieve answers from a large-scale corpus of published COVID-19 scientific papers. Furthermore, we incorporate effective re-ranking and question-answering techniques, such as document diversity and multiple answer spans. Our open-domain question-answering system can further act as a model for the quick development of similar systems that can be adapted and modified for other developing emergent domains.

5

Main Conference: Monday, November 8

Overview

Session 5

09:00 – 10:30	Question Answering 1	Generation 2	Dialogue and Interactive Systems 3	Sentiment Analysis, Stylistic Analysis, and Argument Mining 2	Resources and Evaluation 2	Interpretability and Analysis of Models for NLP 2
	Bavaro 1	*Bavaro 2*	*Bavaro 3*	*Punta Cana 1,2*	*Punta Cana 3,4*	*Santo Domingo*
	[Posters&Demos]: In-person Poster 3					*Bavaro 4*

10:30 – 11:00 **Coffee Break**

11:00 – 12:00 **Invited Talk 2 by Evelina Fedorenko** *Bavaro 1, 2, 3*

12:30 – 14:30 **Virtual Poster Session 2**

14:30 – 14:45 **Mini Break**

Session 6

14:45 – 16:15	Machine Learning for NLP 4	Summarization 2	Machine Translation and Multilinguality 2	Speech, Vision, Robotics, Multimodal Grounding 2	Sentiment Analysis, Stylistic Analysis, and Argument Mining 3	Semantics 2
	Bavaro 1	*Bavaro 2*	*Bavaro 3*	*Punta Cana 1,2*	*Punta Cana 3,4*	*Santo Domingo*
	[Posters&Demos]: In-person Poster 4					*Bavaro 4*

16:15 – 16:45 **Coffee Break**

Session 7

16:45 – 18:15	Machine Translation and Multilinguality 3	Question Answering 2	Dialogue and Interactive Systems 4	Resources and Evaluation 3	Efficient Methods for NLP 2	Semantics 3
	Bavaro 1	*Bavaro 2*	*Bavaro 3*	*Punta Cana 1,2*	*Punta Cana 3,4*	*Santo Domingo*

Keynote Address: Evelina Fedorenko

The Language System in the Human Brain.

Monday, November 8, 2021, 11:00–12:00am

Bavaro 1, 2, 3

Abstract: The goal of my research program is to understand the representations and computations that enable us to share complex thoughts with one another via language, and their neural implementation. A decade ago, I developed a robust new approach to the study of language in the brain based on identifying language-responsive cortex functionally in individual participants. Originally developed for fMRI, we have since extended this approach to other modalities, like intracranial recordings. Using this functional-localization approach, I identified and characterized a set of frontal and temporal brain areas that i) support language comprehension and production (spoken and written); ii) are robustly separable from the lower-level perceptual (e.g., speech processing) and motor (e.g., articulation) brain areas; iii) are spatially and functionally similar across diverse languages (>40 languages from 11 language families); and iv) form a functionally integrated system with substantial redundancy across different components. In this talk, I will highlight a few discoveries from the last decade and argue that the primary goal of language is efficient information transfer rather than enabling complex thought, as has been argued in one prominent philosophical and linguistic tradition (e.g., Wittgenstein, 1921; Berwick Chomsky, 2016). I will use two kinds of evidence to make this argument. First, I will examine the relationship between language and other aspects of cognition, including social cognitive abilities and complex thought/reasoning. I will show that the language brain regions are highly selective for language over diverse non-linguistic processes while also showing a deep and intriguing link with a system that supports social cognition. And second, I will examine different properties of language and argue that language both has a) properties that make it well-suited for communication, and b) properties that make it not suitable for complex thought. Both of these lines of evidence support the communicative function of language, and suggest that the idea that language evolved to allow for more complexity in thought is unlikely.

Biography: Dr. Fedorenko is a cognitive neuroscientist who studies the human language system. She received her bachelor's degree from Harvard University in 2002, and her Ph.D. from the Massachusetts Institute of Technology in 2007. She was then awarded a K99R00 career development award from the National Institute for Child Health and Human Development at the U.S. National Institutes of Health. In 2014, she joined the faculty at Harvard Medical School/Massachusetts General Hospital in Boston, and in 2019 she returned to MIT where she is currently the Frederick A. (1971) and Carole J. Middleton Career Development Associate Professor of Neuroscience in the Brain and Cognitive Sciences Department and the McGovern Institute for Brain Research. Dr. Fedorenko uses fMRI, intracranial recordings and stimulation, EEG/ERPs, MEG, as well as computational modeling, to study adults and children, including those with developmental and acquired brain disorders.

Session 5 Overview – Monday, November 8, 2021

Track A	Track B	Track C	Track D	Track E	Track F	
Question Answering 1	*Generation 2*	*Dialogue and Interactive Systems 3*	*Sentiment Analysis, Stylistic Analysis, and Argument Mining 2*	*Resources and Evaluation 2*	*Interpretability and Analysis of Models for NLP 2*	
Improving Unsupervised Question Answering via Summarization-Informed Question Generation *Lyu et al.*	DiscoDVT: Generating Long Text with Discourse-Aware Discrete Variational Transformer *Ji and Huang*	Maintaining Common Ground in Dynamic Environments *Udagawa and Aizawa*	Dimensional Emotion Detection from Categorical Emotion *Park et al.*	CSDS: A Fine-Grained Chinese Dataset for Customer Service Dialogue Summarization *Lin et al.*	Multi-granularity Textual Adversarial Attack with Behavior Cloning *Chen, Su, and Wei*	9:00
TransferNet: An Effective and Transparent Framework for Multi-hop Question Answering over Relation Graph *Shi et al.*	Mathematical Word Problem Generation from Commonsense Knowledge Graph and Equations *Liu et al.*	Contextualize Knowledge Bases with Transformer for End-to-end Task-Oriented Dialogue Systems *Gou et al.*	Not All Negatives are Equal: Label-Aware Contrastive Loss for Fine-grained Text Classification *Suresh and Ong*	CodRED: A Cross-Document Relation Extraction Dataset for Acquiring Knowledge in the Wild *Yao et al.*	All Bark and No Bite: Rogue Dimensions in Transformer Language Models Obscure Representational Quality *Timkey and Schijndel*	9:15
Topic Transferable Table Question Answering *Chemmengath et al.*	Generic resources are what you need: Style transfer tasks without task-specific parallel training data *Lai, Toral, and Nissim*	Efficient Dialogue Complementary Policy Learning via Deep Q-network Policy and Episodic Memory Policy *Zhao et al.*	Joint Multimodal Aspect-Sentiment Analysis with Auxiliary Cross-modal Relation Detection *Ju et al.*	Building and Evaluating Open-Domain Dialogue Corpora with Clarifying Questions *Ailannejadi et al.*	Incorporating Residual and Normalization Layers into Analysis of Masked Language Models *Kobayashi et al.*	9:30
WebSRC: A Dataset for Web-Based Structural Reading Comprehension *Chen et al.*	Revisiting Pivot-Based Paraphrase Generation: Language Is Not the Only Optional Pivot *Cai, Cao, and Wan*	CRFR: Improving Conversational Recommender Systems via Flexible Fragments Reasoning on Knowledge Graphs *Zhou et al.*	Solving Aspect Category Sentiment Analysis as a Text Generation Task *Liu et al.*	The (Un)Suitability of Automatic Evaluation Metrics for Text Simplification *Alva-Manchego, Scarton, and Specia*	Mind the Style of Text! Adversarial and Backdoor Attacks Based on Text Style Transfer *Qi et al.*	9:45

	Track A	**Track B**	**Track C**	**Track D**	**Track E**	**Track F**
	Question Answering 1	*Generation 2*	*Dialogue and Interactive Systems 3*	*Sentiment Analysis, Stylistic Analysis, and Argument Mining 2*	*Resources and Evaluation 2*	*Interpretability and Analysis of Models for NLP 2*
10:00	Cryptonite: A Cryptic Crossword Benchmark for Extreme Ambiguity in Language *Efrat et al.*	Structural Adapters in Pretrained Language Models for AMR-to-Text Generation *Ribeiro, Zhang, and Gurevych*	DuRecDial 2.0: A Bilingual Parallel Corpus for Conversational Recommendation *Liu et al.*	Semantics-Preserved Data Augmentation for Aspect-Based Sentiment Analysis *Hsu et al.*	We Need to Talk About train-dev-test Splits *Goot*	Sociolectal Analysis of Pretrained Language Models *Zhang et al.*
10:10	End-to-End Entity Resolution and Question Answering Using Differentiable Knowledge Graphs *Saffari et al.*	Data-to-text Generation by Splicing Together Nearest Neighbors *Wiseman, Backurs, and Stratos*	End-to-End Learning of Flowchart Grounded Task-Oriented Dialogs *Raghu, Agarwal, and Joshi*	The Effect of Round-Trip Translation on Fairness in Sentiment Analysis *Christiansen, Gammelgaard, and Søgaard*	PhoMT: A High-Quality and Large-Scale Benchmark Dataset for Vietnamese-English Machine Translation *Doan et al.*	Examining Cross-lingual Contextual Embeddings with Orthogonal Structural Probes *Limisiewicz and Mareček*
10:20	Improving Query Graph Generation for Complex Question Answering over Knowledge Base *Qin et al.*			CHoRaL: Collecting Humor Reaction Labels from Millions of Social Media Users *Yang, Hooshmand, and Hirschberg*	Lying Through One's Teeth: A Study on Verbal Leakage Cues *Yeh and Ku*	Are Transformers a Modern Version of ELIZA? Observations on French Object Verb Agreement *Li, Wisniewski, and Crabbé*

Parallel Session 5

Session 5A: Question Answering 1

Bavaro 1 *Chair: Eunsol Choi*

Improving Unsupervised Question Answering via Summarization-Informed Question Generation

Chenyang Lyu et al. 09:00–9:15

Question Generation (QG) is the task of generating a plausible question for a given <passage, answer> pair. Template-based QG uses linguistically-informed heuristics to transform declarative sentences into interrogatives, whereas supervised QG uses existing Question Answering (QA) datasets to train a system to generate a question given a passage and an answer. A disadvantage of the heuristic approach is that the generated questions are heavily tied to their declarative counterparts. A disadvantage of the supervised approach is that they are heavily tied to the domain/language of the QA dataset used as training data. In order to overcome these shortcomings, we propose a distantly-supervised QG method which uses questions generated heuristically from summaries as a source of training data for a QG system. We make use of freely available news summary data, transforming declarative summary sentences into appropriate questions using heuristics informed by dependency parsing, named entity recognition and semantic role labeling. The resulting questions are then combined with the original news articles to train an end-to-end neural QG model. We extrinsically evaluate our approach using unsupervised QA: our QG model is used to generate synthetic QA pairs for training a QA model. Experimental results show that, trained with only 20k English Wikipedia-based synthetic QA pairs, the QA model substantially outperforms previous unsupervised models on three in-domain datasets (SQuAD1.1, Natural Questions, TriviaQA) and three out-of-domain datasets (NewsQA, BioASQ, DuoRC), demonstrating the transferability of the approach.

TransferNet: An Effective and Transparent Framework for Multi-hop Question Answering over Relation Graph

Jiaxin Shi et al. 09:15–9:30

Multi-hop Question Answering (QA) is a challenging task because it requires precise reasoning with entity relations at every step towards the answer. The relations can be represented in terms of labels in knowledge graph (e.g., spouse) or text in text corpus (e.g., they have been married for 26 years). Existing models usually infer the answer by predicting the sequential relation path or aggregating the hidden graph features. The former is hard to optimize, and the latter lacks interpretability. In this paper, we propose TransferNet, an effective and transparent model for multi-hop QA, which supports both label and text relations in a unified framework. TransferNet jumps across entities at multiple steps. At each step, it attends to different parts of the question, computes activated scores for relations, and then transfer the previous entity scores along activated relations in a differentiable way. We carry out extensive experiments on three datasets and demonstrate that TransferNet surpasses the state-of-the-art models by a large margin. In particular, on MetaQA, it achieves 100% accuracy in 2-hop and 3-hop questions. By qualitative analysis, we show that TransferNet has transparent and interpretable intermediate results.

Topic Transferable Table Question Answering

Saneem Chemmengath et al. 09:30–9:45

Weakly-supervised table question-answering (TableQA) models have achieved state-of-art performance by using pre-trained BERT transformer to jointly encoding a question and a table to produce structured query for the question. However, in practical settings TableQA systems are deployed over table corpora having topic and word distributions quite distinct from BERT's pretraining corpus. In this work we simulate the practical topic shift scenario by designing novel challenge benchmarks WikiSQL-TS and WikiTable-TS, consisting of train-dev-test splits in five distinct topic groups, based on the popular WikiSQL and WikiTable-Questions datasets. We empirically show that, despite pre-training on large open-domain text, performance of models degrades significantly when they are evaluated on unseen topics. In response, we propose T3QA (Topic Transferable Table Question Answering) a pragmatic adaptation framework for TableQA comprising of: (1) topic-specific vocabulary injection into BERT, (2) a novel text-to-text transformer generator (such as T5, GPT2) based natural language question generation pipeline focused on generating topic-specific training data, and (3) a logical form re-ranker. We show that T3QA provides a reasonably good baseline for our topic shift benchmarks. We believe our topic split benchmarks will lead to robust TableQA solutions that are better suited for practical deployment

WebSRC: A Dataset for Web-Based Structural Reading Comprehension

Xingyu Chen et al. 09:45–10:00

Web search is an essential way for humans to obtain information, but it's still a great challenge for machines to understand the contents of web pages. In this paper, we introduce the task of web-based structural reading comprehension. Given a web page and a question about it, the task is to find an answer from the web page. This task requires a system not only to understand the semantics of texts but also the structure of the web page. Moreover, we proposed WebSRC, a novel Web-based Structural Reading Comprehension dataset. WebSRC consists of 400K question-answer pairs, which are collected from 6.4K web pages with corresponding HTML source code, screenshots, and metadata. Each question in WebSRC requires a certain structural understanding of a web page to answer, and the answer is either a text span on the web page or yes/no. We evaluate various strong baselines on our dataset to show the difficulty of our task. We also investigate the usefulness of structural

information and visual features. Our dataset and baselines have been publicly available.

Cryptonite: A Cryptic Crossword Benchmark for Extreme Ambiguity in Language
Avia Efrat et al. 10:00–10:10

Current NLP datasets targeting ambiguity can be solved by a native speaker with relative ease. We present Cryptonite, a large-scale dataset based on cryptic crosswords, which is both linguistically complex and naturally sourced. Each example in Cryptonite is a cryptic clue, a short phrase or sentence with a misleading surface reading, whose solving requires disambiguating semantic, syntactic, and phonetic wordplays, as well as world knowledge. Cryptic clues pose a challenge even for experienced solvers, though top-tier experts can solve them with almost 100% accuracy. Cryptonite is a challenging task for current models; fine-tuning T5-Large on 470k cryptic clues achieves only 7.6% accuracy, on par with the accuracy of a rule-based clue solver (8.6%).

End-to-End Entity Resolution and Question Answering Using Differentiable Knowledge Graphs
Amir Saffari et al. 10:10–10:20

Recently, end-to-end (E2E) trained models for question answering over knowledge graphs (KGQA) have delivered promising results using only a weakly supervised dataset. However, these models are trained and evaluated in a setting where hand-annotated question entities are supplied to the model, leaving the important and non-trivial task of entity resolution (ER) outside the scope of E2E learning. In this work, we extend the boundaries of E2E learning for KGQA to include the training of an ER component. Our model only needs the question text and the answer entities to train, and delivers a stand-alone QA model that does not require an additional ER component to be supplied during runtime. Our approach is fully differentiable, thanks to its reliance on a recent method for building differentiable KGs (Cohen et al., 2020). We evaluate our E2E trained model on two public datasets and show that it comes close to baseline models that use hand-annotated entities.

Improving Query Graph Generation for Complex Question Answering over Knowledge Base
Kechen Qin et al. 10:20–10:30

Most of the existing Knowledge-based Question Answering (KBQA) methods first learn to map the given question to a query graph, and then convert the graph to an executable query to find the answer. The query graph is typically expanded progressively from the topic entity based on a sequence prediction model. In this paper, we propose a new solution to query graph generation that works in the opposite manner: we start with the entire knowledge base and gradually shrink it to the desired query graph. This approach improves both the efficiency and the accuracy of query graph generation, especially for complex multi-hop questions. Experimental results show that our method achieves state-of-the-art performance on ComplexWebQuestion (CWQ) dataset.

Session 5B: Generation 2

Bavaro 2 *Chair: Meng Jiang*

DiscoDVT: Generating Long Text with Discourse-Aware Discrete Variational Transformer

Haozhe Ji and Minlie Huang 09:00–9:15

Despite the recent advances in applying pre-trained language models to generate high-quality texts, generating long passages that maintain long-range coherence is yet challenging for these models. In this paper, we propose DiscoDVT, a discourse-aware discrete variational Transformer to tackle the incoherence issue. DiscoDVT learns a discrete variable sequence that summarizes the global structure of the text and then applies it to guide the generation process at each decoding step. To further embed discourse-aware information into the discrete latent representations, we introduce an auxiliary objective to model the discourse relations within the text. We conduct extensive experiments on two open story generation datasets and demonstrate that the latent codes learn meaningful correspondence to the discourse structures that guide the model to generate long texts with better long-range coherence.

Mathematical Word Problem Generation from Commonsense Knowledge Graph and Equations

Tianqiao Liu et al. 09:15–9:30

There is an increasing interest in the use of mathematical word problem (MWP) generation in educational assessment. Different from standard natural question generation, MWP generation needs to maintain the underlying mathematical operations between quantities and variables, while at the same time ensuring the relevance between the output and the given topic. To address above problem, we develop an end-to-end neural model to generate diverse MWPs in real-world scenarios from commonsense knowledge graph and equations. The proposed model (1) learns both representations from edge-enhanced Levi graphs of symbolic equations and commonsense knowledge; (2) automatically fuses equation and commonsense knowledge information via a self-planning module when generating the MWPs. Experiments on an educational gold-standard set and a large-scale generated MWP set show that our approach is superior on the MWP generation task, and it outperforms the SOTA models in terms of both automatic evaluation metrics, i.e., BLEU-4, ROUGE-L, Self-BLEU, and human evaluation metrics, i.e., equation relevance, topic relevance, and language coherence. To encourage reproducible results, we make our code and MWP dataset public available at `https://github.com/tal-ai/MaKE_EMNLP2021`.

Generic resources are what you need: Style transfer tasks without task-specific parallel training data

Huiyuan Lai, Antonio Toral, and Malvina Nissim 09:30–9:45

Style transfer aims to rewrite a source text in a different target style while preserving its content. We propose a novel approach to this task that leverages generic resources, and without using any task-specific parallel (source—target) data outperforms existing unsupervised approaches on the two most popular style transfer tasks: formality transfer and polarity swap. In practice, we adopt a multi-step procedure which builds on a generic pre-trained sequence-to-sequence model (BART). First, we strengthen the model's ability to rewrite by further pre-training BART on both an existing collection of generic paraphrases, as well as on synthetic pairs created using a general-purpose lexical resource. Second, through an iterative back-translation approach, we train two models, each in a transfer direction, so that they can provide each other with synthetically generated pairs, dynamically in the training process. Lastly, we let our best resulting model generate static synthetic pairs to be used in a supervised training regime. Besides methodology and state-of-the-art results, a core contribution of this work is a reflection on the nature of the two tasks we address, and how their differences are highlighted by their response to our approach.

Revisiting Pivot-Based Paraphrase Generation: Language Is Not the Only Optional Pivot

Yitao Cai, Yue Cao, and Xiaojun Wan 09:45–10:00

Paraphrases refer to texts that convey the same meaning with different expression forms. Pivot-based methods, also known as the round-trip translation, have shown promising results in generating high-quality paraphrases. However, existing pivot-based methods all rely on language as the pivot, where large-scale, high-quality parallel bilingual texts are required. In this paper, we explore the feasibility of using semantic and syntactic representations as the pivot for paraphrase generation. Concretely, we transform a sentence into a variety of different semantic or syntactic representations (including AMR, UD, and latent semantic representation), and then decode the sentence back from the semantic representations. We further explore a pretraining-based approach to compress the pipeline process into an end-to-end framework. We conduct experiments comparing different approaches with different kinds of pivots. Experimental results show that taking AMR as pivot can obtain paraphrases with better quality than taking language as the pivot. The end-to-end framework can reduce semantic shift when language is used as the pivot. Besides, several unsupervised pivot-based methods can generate paraphrases with similar quality as the supervised sequence-to-sequence model, which indicates that parallel data of paraphrases may not be necessary for paraphrase generation.

Structural Adapters in Pretrained Language Models for AMR-to-Text Generation

Leonardo F. R. Ribeiro, Yue Zhang, and Iryna Gurevych 10:00–10:15

Pretrained language models (PLM) have recently advanced graph-to-text generation, where the input graph is linearized into a sequence and fed into the PLM to obtain its representation. However, efficiently encoding

the graph structure in PLMs is challenging because such models were pretrained on natural language, and modeling structured data may lead to catastrophic forgetting of distributional knowledge. In this paper, we propose StructAdapt, an adapter method to encode graph structure into PLMs. Contrary to prior work, StructAdapt effectively models interactions among the nodes based on the graph connectivity, only training graph structure-aware adapter parameters. In this way, we incorporate task-specific knowledge while maintaining the topological structure of the graph. We empirically show the benefits of explicitly encoding graph structure into PLMs using StructAdapt, outperforming the state of the art on two AMR-to-text datasets, training only 5.1% of the PLM parameters.

Data-to-text Generation by Splicing Together Nearest Neighbors
Sam Wiseman, Arturs Backurs, and Karl Stratos 10:15–10:30

We propose to tackle data-to-text generation tasks by directly splicing together retrieved segments of text from "neighbor" source-target pairs. Unlike recent work that conditions on retrieved neighbors but generates text token-by-token, left-to-right, we learn a policy that directly manipulates segments of neighbor text, by inserting or replacing them in partially constructed generations. Standard techniques for training such a policy require an oracle derivation for each generation, and we prove that finding the shortest such derivation can be reduced to parsing under a particular weighted context-free grammar. We find that policies learned in this way perform on par with strong baselines in terms of automatic and human evaluation, but allow for more interpretable and controllable generation.

Session 5C: Dialogue and Interactive Systems 3

Bavaro 3 *Chair: Zhiyang Teng*

Maintaining Common Ground in Dynamic Environments

Takuma Udagawa and Akiko Aizawa 09:00–9:15

Common grounding is the process of creating and maintaining mutual understandings, which is a critical aspect of sophisticated human communication. While various task settings have been proposed in existing literature, they mostly focus on creating common ground under static context and ignore the aspect of maintaining them overtime under dynamic context. In this work, we propose a novel task setting to study the ability of both creating and maintaining common ground in dynamic environments. Based on our minimal task formulation, we collected a large-scale dataset of 5,617 dialogues to enable fine-grained evaluation and analysis of various dialogue systems. Through our dataset analyses, we highlight novel challenges introduced in our setting, such as the usage of complex spatio-temporal expressions to create and maintain common ground. Finally, we conduct extensive experiments to assess the capabilities of our baseline dialogue system and discuss future prospects of our research.

Contextualize Knowledge Bases with Transformer for End-to-end Task-Oriented Dialogue Systems

Yanjie Gou et al. 09:15–9:30

Incorporating knowledge bases (KB) into end-to-end task-oriented dialogue systems is challenging, since it requires to properly represent the entity of KB, which is associated with its KB context and dialogue context. The existing works represent the entity with only perceiving a part of its KB context, which can lead to the less effective representation due to the information loss, and adversely favor KB reasoning and response generation. To tackle this issue, we explore to fully contextualize the entity representation by dynamically perceiving all the relevant entities and dialogue history. To achieve this, we propose a COntext-aware Memory Enhanced Transformer framework (COMET), which treats the KB as a sequence and leverages a novel Memory Mask to enforce the entity to only focus on its relevant entities and dialogue history, while avoiding the distraction from the irrelevant entities. Through extensive experiments, we show that our COMET framework can achieve superior performance over the state of the arts.

Efficient Dialogue Complementary Policy Learning via Deep Q-network Policy and Episodic Memory Policy

Yangyang Zhao et al. 09:30–9:45

Deep reinforcement learning has shown great potential in training dialogue policies. However, its favorable performance comes at the cost of many rounds of interaction. Most of the existing dialogue policy methods rely on a single learning system, while the human brain has two specialized learning and memory systems, supporting to find good solutions without requiring copious examples. Inspired by the human brain, this paper proposes a novel complementary policy learning (CPL) framework, which exploits the complementary advantages of the episodic memory (EM) policy and the deep Q-network (DQN) policy to achieve fast and effective dialogue policy learning. In order to coordinate between the two policies, we proposed a confidence controller to control the complementary time according to their relative efficacy at different stages. Furthermore, memory connectivity and time pruning are proposed to guarantee the flexible and adaptive generalization of the EM policy in dialog tasks. Experimental results on three dialogue datasets show that our method significantly outperforms existing methods relying on a single learning system.

CRFR: Improving Conversational Recommender Systems via Flexible Fragments Reasoning on Knowledge Graphs

Jinfeng Zhou et al. 09:45–10:00

Although paths of user interests shift in knowledge graphs (KGs) can benefit conversational recommender systems (CRS), explicit reasoning on KGs has not been well considered in CRS, due to the complex of high-order and incomplete paths. We propose CRFR, which effectively does explicit multi-hop reasoning on KGs with a conversational context-based reinforcement learning model. Considering the incompleteness of KGs, instead of learning single complete reasoning path, CRFR flexibly learns multiple reasoning fragments which are likely contained in the complete paths of interests shift. A fragments-aware unified model is then designed to fuse the fragments information from item-oriented and concept-oriented KGs to enhance the CRS response with entities and words from the fragments. Extensive experiments demonstrate CRFR's SOTA performance on recommendation, conversation and conversation interpretability.

DuRecDial 2.0: A Bilingual Parallel Corpus for Conversational Recommendation

Zeming Liu et al. 10:00–10:15

In this paper, we provide a bilingual parallel human-to-human recommendation dialog dataset (DuRecDial 2.0) to enable researchers to explore a challenging task of multilingual and cross-lingual conversational recommendation. The difference between DuRecDial 2.0 and existing conversational recommendation datasets is that the data item (Profile, Goal, Knowledge, Context, Response) in DuRecDial 2.0 is annotated in two languages, both English and Chinese, while other datasets are built with the setting of a single language. We collect 8.2k dialogs aligned across English and Chinese languages (16.5k dialogs and 255k utterances in total) that are annotated by crowdsourced workers with strict quality control procedure. We then build monolingual, mul-

tilingual, and cross-lingual conversational recommendation baselines on DuRecDial 2.0. Experiment results show that the use of additional English data can bring performance improvement for Chinese conversational recommendation, indicating the benefits of DuRecDial 2.0. Finally, this dataset provides a challenging testbed for future studies of monolingual, multilingual, and cross-lingual conversational recommendation.

End-to-End Learning of Flowchart Grounded Task-Oriented Dialogs
Dinesh Raghu, Shantanu Agarwal, and Sachindra Joshi 10:15–10:30

We propose a novel problem within end-to-end learning of task oriented dialogs (TOD), in which the dialog system mimics a troubleshooting agent who helps a user by diagnosing their problem (e.g., car not starting). Such dialogs are grounded in domain-specific flowcharts, which the agent is supposed to follow during the conversation. Our task exposes novel technical challenges for neural TOD, such as grounding an utterance to the flowchart without explicit annotation, referring to additional manual pages when user asks a clarification question, and ability to follow unseen flowcharts at test time. We release a dataset (FLODIAL) consisting of 2,738 dialogs grounded on 12 different troubleshooting flowcharts. We also design a neural model, FLONET, which uses a retrieval-augmented generation architecture to train the dialog agent. Our experiments find that FLONET can do zero-shot transfer to unseen flowcharts, and sets a strong baseline for future research.

Session 5D: Sentiment Analysis, Stylistic Analysis, and Argument Mining 2

Punta Cana 1,2 *Chair: Veronique Hoste*

Dimensional Emotion Detection from Categorical Emotion

Sungjoon Park et al. 09:00–9:15

We present a model to predict fine-grained emotions along the continuous dimensions of valence, arousal, and dominance (VAD) with a corpus with categorical emotion annotations. Our model is trained by minimizing the EMD (Earth Mover's Distance) loss between the predicted VAD score distribution and the categorical emotion distributions sorted along VAD, and it can simultaneously classify the emotion categories and predict the VAD scores for a given sentence. We use pre-trained RoBERTa-Large and fine-tune on three different corpora with categorical labels and evaluate on EmoBank corpus with VAD scores. We show that our approach reaches comparable performance to that of the state-of-the-art classifiers in categorical emotion classification and shows significant positive correlations with the ground truth VAD scores. Also, further training with supervision of VAD labels leads to improved performance especially when dataset is small. We also present examples of predictions of appropriate emotion words that are not part of the original annotations.

Not All Negatives are Equal: Label-Aware Contrastive Loss for Fine-grained Text Classification

Varsha Suresh and Desmond Ong 09:15–9:30

Fine-grained classification involves dealing with datasets with larger number of classes with subtle differences between them. Guiding the model to focus on differentiating dimensions between these commonly confusable classes is key to improving performance on fine-grained tasks. In this work, we analyse the contrastive fine-tuning of pre-trained language models on two fine-grained text classification tasks, emotion classification and sentiment analysis. We adaptively embed class relationships into a contrastive objective function to help differently weigh the positives and negatives, and in particular, weighting closely confusable negatives more than less similar negative examples. We find that Label-aware Contrastive Loss outperforms previous contrastive methods, in the presence of larger number and/or more confusable classes, and helps models to produce output distributions that are more differentiated.

Joint Multi-modal Aspect-Sentiment Analysis with Auxiliary Cross-modal Relation Detection

Xincheng Ju et al. 09:30–9:45

Aspect terms extraction (ATE) and aspect sentiment classification (ASC) are two fundamental and fine-grained sub-tasks in aspect-level sentiment analysis (ALSA). In the textual analysis, joint extracting both aspect terms and sentiment polarities has been drawn much attention due to the better applications than individual sub-task. However, in the multi-modal scenario, the existing studies are limited to handle each sub-task independently, which fails to model the innate connection between the above two objectives and ignores the better applications. Therefore, in this paper, we are the first to jointly perform multi-modal ATE (MATE) and multi-modal ASC (MASC), and we propose a multi-modal joint learning approach with auxiliary cross-modal relation detection for multi-modal aspect-level sentiment analysis (MALSA). Specifically, we first build an auxiliary text-image relation detection module to control the proper exploitation of visual information. Second, we adopt the hierarchical framework to bridge the multi-modal connection between MATE and MASC, as well as separately visual guiding for each sub module. Finally, we can obtain all aspect-level sentiment polarities dependent on the jointly extracted specific aspects. Extensive experiments show the effectiveness of our approach against the joint textual approaches, pipeline and collapsed multi-modal approaches.

Solving Aspect Category Sentiment Analysis as a Text Generation Task

Jian Liu et al. 09:45–10:00

Aspect category sentiment analysis has attracted increasing research attention. The dominant methods make use of pre-trained language models by learning effective aspect category-specific representations, and adding specific output layers to its pre-trained representation. We consider a more direct way of making use of pre-trained language models, by casting the ACSA tasks into natural language generation tasks, using natural language sentences to represent the output. Our method allows more direct use of pre-trained knowledge in seq2seq language models by directly following the task setting during pre-training. Experiments on several benchmarks show that our method gives the best reported results, having large advantages in few-shot and zero-shot settings.

Semantics-Preserved Data Augmentation for Aspect-Based Sentiment Analysis

Ting-Wei Hsu et al. 10:00–10:10

Both the issues of data deficiencies and semantic consistency are important for data augmentation. Most of previous methods address the first issue, but ignore the second one. In the cases of aspect-based sentiment analysis, violation of the above issues may change the aspect and sentiment polarity. In this paper, we propose a semantics-preservation data augmentation approach by considering the importance of each word in a textual sequence according to the related aspects and sentiments. We then substitute the unimportant tokens with two replacement strategies without altering the aspect-level polarity. Our approach is evaluated on several publicly available sentiment analysis datasets and the real-world stock price/risk movement prediction scenarios. Experimental results show that our methodology achieves better performances in all datasets.

The Effect of Round-Trip Translation on Fairness in Sentiment Analysis
Jonathan Christiansen, Mathias Gammelgaard, and Anders Søgaard 10:10–10:20

Sentiment analysis systems have been shown to exhibit sensitivity to protected attributes. Round-trip translation, on the other hand, has been shown to normalize text. We explore the impact of round-trip translation on the demographic parity of sentiment classifiers and show how round-trip translation consistently improves classification fairness at test time (reducing up to 47% of between-group gaps). We also explore the idea of retraining sentiment classifiers on round-trip-translated data.

CHoRaL: Collecting Humor Reaction Labels from Millions of Social Media Users
Zixiaofan Yang, Shayan Hooshmand, and Julia Hirschberg 10:20–10:30

Humor detection has gained attention in recent years due to the desire to understand user-generated content with figurative language. However, substantial individual and cultural differences in humor perception make it very difficult to collect a large-scale humor dataset with reliable humor labels. We propose CHoRaL, a framework to generate perceived humor labels on Facebook posts, using the naturally available user reactions to these posts with no manual annotation needed. CHoRaL provides both binary labels and continuous scores of humor and non-humor. We present the largest dataset to date with labeled humor on 785K posts related to COVID-19. Additionally, we analyze the expression of COVID-related humor in social media by extracting lexico-semantic and affective features from the posts, and build humor detection models with performance similar to humans. CHoRaL enables the development of large-scale humor detection models on any topic and opens a new path to the study of humor on social media.

Session 5E: Resources and Evaluation 2

Punta Cana 3,4 *Chair: Annemarie Friedrich*

CSDS: A Fine-Grained Chinese Dataset for Customer Service Dialogue Summarization

Haitao Lin et al. 09:00–9:15

Dialogue summarization has drawn much attention recently. Especially in the customer service domain, agents could use dialogue summaries to help boost their works by quickly knowing customer's issues and service progress. These applications require summaries to contain the perspective of a single speaker and have a clear topic flow structure, while neither are available in existing datasets. Therefore, in this paper, we introduce a novel Chinese dataset for Customer Service Dialogue Summarization (CSDS). CSDS improves the abstractive summaries in two aspects: (1) In addition to the overall summary for the whole dialogue, role-oriented summaries are also provided to acquire different speakers' viewpoints. (2) All the summaries sum up each topic separately, thus containing the topic-level structure of the dialogue. We define tasks in CSDS as generating the overall summary and different role-oriented summaries for a given dialogue. Next, we compare various summarization methods on CSDS, and experiment results show that existing methods are prone to generate redundant and incoherent summaries. Besides, the performance becomes much worse when analyzing the performance on role-oriented summaries and topic structures. We hope that this study could benchmark Chinese dialogue summarization and benefit further studies.

CodRED: A Cross-Document Relation Extraction Dataset for Acquiring Knowledge in the Wild

Yuan Yao et al. 09:15–9:30

Existing relation extraction (RE) methods typically focus on extracting relational facts between entity pairs within single sentences or documents. However, a large quantity of relational facts in knowledge bases can only be inferred across documents in practice. In this work, we present the problem of cross-document RE, making an initial step towards knowledge acquisition in the wild. To facilitate the research, we construct the first human-annotated cross-document RE dataset CodRED. Compared to existing RE datasets, CodRED presents two key challenges: Given two entities, (1) it requires finding the relevant documents that can provide clues for identifying their relations; (2) it requires reasoning over multiple documents to extract the relational facts. We conduct comprehensive experiments to show that CodRED is challenging to existing RE methods including strong BERT-based models.

Building and Evaluating Open-Domain Dialogue Corpora with Clarifying Questions

Mohammad Ailannejadi et al. 09:30–9:45

Enabling open-domain dialogue systems to ask clarifying questions when appropriate is an important direction for improving the quality of the system response. Namely, for cases when a user request is not specific enough for a conversation system to provide an answer right away, it is desirable to ask a clarifying question to increase the chances of retrieving a satisfying answer. To address the problem of 'asking clarifying questions in open-domain dialogues': (1) we collect and release a new dataset focused on open-domain single- and multi-turn conversations, (2) we benchmark several state-of-the-art neural baselines, and (3) we propose a pipeline consisting of offline and online steps for evaluating the quality of clarifying questions in various dialogues. These contributions are suitable as a foundation for further research.

The (Un)Suitability of Automatic Evaluation Metrics for Text Simplification

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia 09:45–10:00

In order to simplify sentences, several rewriting operations can be performed, such as replacing complex words per simpler synonyms, deleting unnecessary information, and splitting long sentences. Despite this multi-operation nature, evaluation of automatic simplification systems relies on metrics that moderately correlate with human judgments on the simplicity achieved by executing specific operations (e.g., simplicity gain based on lexical replacements). In this article, we investigate how well existing metrics can assess sentence-level simplifications where multiple operations may have been applied and which, therefore, require more general simplicity judgments. For that, we first collect a new and more reliable data set for evaluating the correlation of metrics and human judgments of overall simplicity. Second, we conduct the first metaevaluation of automatic metrics in Text Simplification, using our new data set (and other existing data) to analyze the variation of the correlation between metrics' scores and human judgments across three dimensions: the perceived simplicity level, the system type, and the set of references used for computation. We show that these three aspects affect the correlations and, in particular, highlight the limitations of commonly used operation-specific metrics. Finally, based on our findings, we propose a set of recommendations for automatic evaluation of multi-operation simplifications, suggesting which metrics to compute and how to interpret their scores.

We Need to Talk About train-dev-test Splits

Rob van der Goot 10:00–10:10

Standard train-dev-test splits used to benchmark multiple models against each other are ubiquitously used in Natural Language Processing (NLP). In this setup, the train data is used for training the model, the development set for evaluating different versions of the proposed model(s) during development, and the test set to confirm the answers to the main research question(s). However, the introduction of neural networks in NLP has led to a different use of these standard splits; the development set is now often used for model selection during the training procedure. Because of this, comparing multiple versions of the same model during development leads to overestimation on the development data. As an effect, people have started to compare an increasing amount

of models on the test data, leading to faster overfitting and "expiration" of our test sets. We propose to use a tune-set when developing neural network methods, which can be used for model picking so that comparing the different versions of a new model can safely be done on the development data.

PhoMT: A High-Quality and Large-Scale Benchmark Dataset for Vietnamese-English Machine Translation
Long Doan et al. 10:10–10:20

We introduce a high-quality and large-scale Vietnamese-English parallel dataset of 3.02M sentence pairs, which is 2.9M pairs larger than the benchmark Vietnamese-English machine translation corpus IWSLT15. We conduct experiments comparing strong neural baselines and well-known automatic translation engines on our dataset and find that in both automatic and human evaluations: the best performance is obtained by fine-tuning the pre-trained sequence-to-sequence denoising auto-encoder mBART. To our best knowledge, this is the first large-scale Vietnamese-English machine translation study. We hope our publicly available dataset and study can serve as a starting point for future research and applications on Vietnamese-English machine translation. We release our dataset at: https://github.com/VinAIResearch/PhoMT

Lying Through One's Teeth: A Study on Verbal Leakage Cues
Min-Hsuan Yeh and Lun-Wei Ku 10:20–10:30

Although many studies use the LIWC lexicon to show the existence of verbal leakage cues in lie detection datasets, none mention how verbal leakage cues are influenced by means of data collection, or the impact thereof on the performance of models. In this paper, we study verbal leakage cues to understand the effect of the data construction method on their significance, and examine the relationship between such cues and models' validity. The LIWC word-category dominance scores of seven lie detection datasets are used to show that audio statements and lie-based annotations indicate a greater number of strong verbal leakage cue categories. Moreover, we evaluate the validity of state-of-the-art lie detection models with cross- and in-dataset testing. Results show that in both types of testing, models trained on a dataset with more strong verbal leakage cue categories—as opposed to only a greater number of strong cues—yield superior results, suggesting that verbal leakage cues are a key factor for selecting lie detection datasets.

Session 5F: Interpretability and Analysis of Models for NLP 2

Santo Domingo *Chair: Leila Arras*

Multi-granularity Textual Adversarial Attack with Behavior Cloning

Yangyi Chen, Jin Su, and Wei Wei 09:00–9:15

Recently, the textual adversarial attack models become increasingly popular due to their successful in estimating the robustness of NLP models. However, existing works have obvious deficiencies. (1)They usually consider only a single granularity of modification strategies (e.g. word-level or sentence-level), which is insufficient to explore the holistic textual space for generation; (2) They need to query victim models hundreds of times to make a successful attack, which is highly inefficient in practice. To address such problems, in this paper we propose MAYA, a Multi-grAnularitY Attack model to effectively generate high-quality adversarial samples with fewer queries to victim models. Furthermore, we propose a reinforcement-learning based method to train a multi-granularity attack agent through behavior cloning with the expert knowledge from our MAYA algorithm to further reduce the query times. Additionally, we also adapt the agent to attack black-box models that only output labels without confidence scores. We conduct comprehensive experiments to evaluate our attack models by attacking BiLSTM, BERT and RoBERTa in two different black-box attack settings and three benchmark datasets. Experimental results show that our models achieve overall better attacking performance and produce more fluent and grammatical adversarial samples compared to baseline models. Besides, our adversarial attack agent significantly reduces the query times in both attack settings. Our codes are released at https://github.com/Yangyi-Chen/MAYA.

All Bark and No Bite: Rogue Dimensions in Transformer Language Models Obscure Representational Quality

William Timkey and Marten van Schijndel 09:15–9:30

Similarity measures are a vital tool for understanding how language models represent and process language. Standard representational similarity measures such as cosine similarity and Euclidean distance have been successfully used in static word embedding models to understand how words cluster in semantic space. Recently, these measures have been applied to embeddings from contextualized models such as BERT and GPT-2. In this work, we call into question the informativity of such measures for contextualized language models. We find that a small number of rogue dimensions, often just 1-3, dominate these measures. Moreover, we find a striking mismatch between the dimensions that dominate similarity measures and those which are important to the behavior of the model. We show that simple postprocessing techniques such as standardization are able to correct for rogue dimensions and reveal underlying representational quality. We argue that accounting for rogue dimensions is essential for any similarity-based analysis of contextual language models.

Incorporating Residual and Normalization Layers into Analysis of Masked Language Models

Goro Kobayashi et al. 09:30–9:45

Transformer architecture has become ubiquitous in the natural language processing field. To interpret the Transformer-based models, their attention patterns have been extensively analyzed. However, the Transformer architecture is not only composed of the multi-head attention; other components can also contribute to Transformers' progressive performance. In this study, we extended the scope of the analysis of Transformers from solely the attention patterns to the whole attention block, i.e., multi-head attention, residual connection, and layer normalization. Our analysis of Transformer-based masked language models shows that the token-to-token interaction performed via attention has less impact on the intermediate representations than previously assumed. These results provide new intuitive explanations of existing reports; for example, discarding the learned attention patterns tends not to adversely affect the performance. The codes of our experiments are publicly available.

Mind the Style of Text! Adversarial and Backdoor Attacks Based on Text Style Transfer

Fanchao Qi et al. 09:45–10:00

Adversarial attacks and backdoor attacks are two common security threats that hang over deep learning. Both of them harness task-irrelevant features of data in their implementation. Text style is a feature that is naturally irrelevant to most NLP tasks, and thus suitable for adversarial and backdoor attacks. In this paper, we make the first attempt to conduct adversarial and backdoor attacks based on text style transfer, which is aimed at altering the style of a sentence while preserving its meaning. We design an adversarial attack method and a backdoor attack method, and conduct extensive experiments to evaluate them. Experimental results show that popular NLP models are vulnerable to both adversarial and backdoor attacks based on text style transfer—the attack success rates can exceed 90% without much effort. It reflects the limited ability of NLP models to handle the feature of text style that has not been widely realized. In addition, the style transfer-based adversarial and backdoor attack methods show superiority to baselines in many aspects. All the code and data of this paper can be obtained at https://github.com/thunlp/StyleAttack.

Sociolectal Analysis of Pretrained Language Models

Sheng Zhang et al. 10:00–10:10

Using data from English cloze tests, in which subjects also self-reported their gender, age, education, and race, we examine performance differences of pretrained language models across demographic groups, defined by these (protected) attributes. We demonstrate wide performance gaps across demographic groups and show

that pretrained language models systematically disfavor young non-white male speakers; i.e., not only do pretrained language models learn social biases (stereotypical associations) – pretrained language models also learn sociolectal biases, learning to speak more like some than like others. We show, however, that, with the exception of BERT models, larger pretrained language models reduce some the performance gaps between majority and minority groups.

Examining Cross-lingual Contextual Embeddings with Orthogonal Structural Probes
Tomasz Limisiewicz and David Mareček 10:10–10:20

State-of-the-art contextual embeddings are obtained from large language models available only for a few languages. For others, we need to learn representations using a multilingual model. There is an ongoing debate on whether multilingual embeddings can be aligned in a space shared across many languages. The novel Orthogonal Structural Probe (Limisiewicz and Mareček, 2021) allows us to answer this question for specific linguistic features and learn a projection based only on mono-lingual annotated datasets. We evaluate syntactic (UD) and lexical (WordNet) structural information encoded inmBERT's contextual representations for nine diverse languages. We observe that for languages closely related to English, no transformation is needed. The evaluated information is encoded in a shared cross-lingual embedding space. For other languages, it is beneficial to apply orthogonal transformation learned separately for each language. We successfully apply our findings to zero-shot and few-shot cross-lingual parsing.

Are Transformers a Modern Version of ELIZA? Observations on French Object Verb Agreement
Bingzhi Li, Guillaume Wisniewski, and Benoit Crabbé 10:20–10:30

Many recent works have demonstrated that unsupervised sentence representations of neural networks encode syntactic information by observing that neural language models are able to predict the agreement between a verb and its subject. We take a critical look at this line of research by showing that it is possible to achieve high accuracy on this agreement task with simple surface heuristics, indicating a possible flaw in our assessment of neural networks' syntactic ability. Our fine-grained analyses of results on the long-range French object-verb agreement show that contrary to LSTMs, Transformers are able to capture a non-trivial amount of grammatical structure.

Poster and Demo session 3

In-person Poster Session 3

Bavaro 4 *09:00–10:30*

Fine-grained Entity Typing via Label Reasoning

Qing Liu et al.

Conventional entity typing approaches are based on independent classification paradigms, which make them difficult to recognize inter-dependent, long-tailed and fine-grained entity types. In this paper, we argue that the implicitly entailed extrinsic and intrinsic dependencies between labels can provide critical knowledge to tackle the above challenges. To this end, we propose Label Reasoning Network(LRN), which sequentially reasons fine-grained entity labels by discovering and exploiting label dependencies knowledge entailed in the data. Specifically, LRN utilizes an auto-regressive network to conduct deductive reasoning and a bipartite attribute graph to conduct inductive reasoning between labels, which can effectively model, learn and reason complex label dependencies in a sequence-to-set, end-to-end manner. Experiments show that LRN achieves the state-of-the-art performance on standard ultra fine-grained entity typing benchmarks, and can also resolve the long tail label problem effectively.

Enhanced Language Representation with Label Knowledge for Span Extraction

Pan Yang et al.

Span extraction, aiming to extract text spans (such as words or phrases) from plain text, is a fundamental process in Information Extraction. Recent works introduce the label knowledge to enhance the text representation by formalizing the span extraction task into a question answering problem (QA Formalization), which achieves state-of-the-art performance. However, such a QA Formalization does not fully exploit the label knowledge and causes a dramatic decrease in efficiency of training/inference. To address those problems, we introduce a fresh paradigm to integrate label knowledge and further propose a novel model to explicitly and efficiently integrate label knowledge into text representations. Specifically, it encodes texts and label annotations independently and then integrates label knowledge into text representation with an elaborate-designed semantics fusion module. We conduct extensive experiments on three typical span extraction tasks: flat NER, nested NER, and event detection. The empirical results show that 1) our model achieves a new state-of-the-art performance on four benchmarks, and 2) reduces training time and inference time by 76% and 77% on average, respectively, compared with the QA Formalization paradigm.

PRIDE: Predicting Relationships in Conversations

Anna Tigunova et al.

Automatically extracting interpersonal relationships of conversation interlocutors can enrich personal knowledge bases to enhance personalized search, recommenders and chatbots. To infer speakers' relationships from dialogues we propose PRIDE, a neural multi-label classifier, based on BERT and Transformer for creating a conversation representation. PRIDE utilizes dialogue structure and augments it with external knowledge about speaker features and conversation style.Unlike prior works, we address multi-label prediction of fine-grained relationships. We release large-scale datasets, based on screenplays of movies and TV shows, with directed relationships of conversation participants. Extensive experiments on both datasets show superior performance of PRIDE compared to the state-of-the-art baselines.

Extracting Fine-Grained Knowledge Graphs of Scientific Claims: Dataset and Transformer-Based Results

Ian Magnusson and Scott Friedman

Recent transformer-based approaches demonstrate promising results on relational scientific information extraction. Existing datasets focus on high-level description of how research is carried out. Instead we focus on the subtleties of how experimental associations are presented by building SciClaim, a dataset of scientific claims drawn from Social and Behavior Science (SBS), PubMed, and CORD-19 papers. Our novel graph annotation schema incorporates not only coarse-grained entity spans as nodes and relations as edges between them, but also fine-grained attributes that modify entities and their relations, for a total of 12,738 labels in the corpus. By including more label types and more than twice the label density of previous datasets, SciClaim captures causal, comparative, predictive, statistical, and proportional associations over experimental variables along with their qualifications, subtypes, and evidence. We extend work in transformer-based joint entity and relation extraction to effectively infer our schema, showing the promise of fine-grained knowledge graphs in scientific claims and beyond.

Sequential Cross-Document Coreference Resolution

Emily Allaway, Shuai Wang, and Miguel Ballesteros

Relating entities and events in text is a key component of natural language understanding. Cross-document coreference resolution, in particular, is important for the growing interest in multi-document analysis tasks. In this work we propose a new model that extends the efficient sequential prediction paradigm for coreference resolution to cross-document settings and achieves competitive results for both entity and event coreference while providing strong evidence of the efficacy of both sequential models and higher-order inference in cross-document settings. Our model incrementally composes mentions into cluster representations and predicts links

between a mention and the already constructed clusters, approximating a higher-order model. In addition, we conduct extensive ablation studies that provide new insights into the importance of various inputs and representation types in coreference.

Mixture-of-Partitions: Infusing Large Biomedical Knowledge Graphs into BERT

Zaiqiao Meng et al.

Infusing factual knowledge into pre-trained models is fundamental for many knowledge-intensive tasks. In this paper, we proposed Mixture-of-Partitions (MoP), an infusion approach that can handle a very large knowledge graph (KG) by partitioning it into smaller sub-graphs and infusing their specific knowledge into various BERT models using lightweight adapters. To leverage the overall factual knowledge for a target task, these sub-graph adapters are further fine-tuned along with the underlying BERT through a mixture layer. We evaluate our MoP with three biomedical BERTs (SciBERT, BioBERT, PubmedBERT) on six downstream tasks (inc. NLI, QA, Classification), and the results show that our MoP consistently enhances the underlying BERTs in task performance, and achieves new SOTA performances on five evaluated datasets.

Filling the Gaps in Ancient Akkadian Texts: A Masked Language Modelling Approach

Koren Lazar et al.

We present models which complete missing text given transliterations of ancient Mesopotamian documents, originally written on cuneiform clay tablets (2500 BCE - 100 CE). Due to the tablets' deterioration, scholars often rely on contextual cues to manually fill in missing parts in the text in a subjective and time-consuming process. We identify that this challenge can be formulated as a masked language modelling task, used mostly as a pretraining objective for contextualized language models. Following, we develop several architectures focusing on the Akkadian language, the lingua franca of the time. We find that despite data scarcity (1M tokens) we can achieve state of the art performance on missing tokens prediction (89% hit5) using a greedy decoding scheme and pretraining on data from other languages and different time periods. Finally, we conduct human evaluations showing the applicability of our models in assisting experts to transcribe texts in extinct languages.

AVocaDo: Strategy for Adapting Vocabulary to Downstream Domain

Jimin Hong et al.

During the fine-tuning phase of transfer learning, the pretrained vocabulary remains unchanged, while model parameters are updated. The vocabulary generated based on the pretrained data is suboptimal for downstream data when domain discrepancy exists. We propose to consider the vocabulary as an optimizable parameter, allowing us to update the vocabulary by expanding it with domain specific vocabulary based on a tokenization statistic. Furthermore, we preserve the embeddings of the added words from overfitting to downstream data by utilizing knowledge learned from a pretrained language model with a regularization term. Our method achieved consistent performance improvements on diverse domains (i.e., biomedical, computer science, news, and reviews).

Can We Improve Model Robustness through Secondary Attribute Counterfactuals?

Ananth Balashankar et al.

Developing robust NLP models that perform well on many, even small, slices of data is a significant but important challenge, with implications from fairness to general reliability. To this end, recent research has explored how models rely on spurious correlations, and how counterfactual data augmentation (CDA) can mitigate such issues. In this paper we study how and why modeling counterfactuals over multiple attributes can go significantly further in improving model performance. We propose RDI, a context-aware methodology which takes into account the impact of secondary attributes on the model's predictions and increases sensitivity for secondary attributes over reweighted counterfactually augmented data. By implementing RDI in the context of toxicity detection, we find that accounting for secondary attributes can significantly improve robustness, with improvements in sliced accuracy on the original dataset up to 7% compared to existing robustness methods. We also demonstrate that RDI generalizes to the coreference resolution task and provide guidelines to extend this to other tasks.

Long-Range Modeling of Source Code Files with eWASH: Extended Window Access by Syntax Hierarchy

Colin Clement et al.

Statistical language modeling and translation with transformers have found many successful applications in program understanding and generation tasks, setting high benchmarks for tools in modern software development environments. The finite context window of these neural models means, however, that they will be unable to leverage the entire relevant context of large files and packages for any given task. While there are many efforts to extend the context window, we introduce an architecture-independent approach for leveraging the syntactic hierarchies of source code for incorporating entire file-level context into a fixed-length window. Using concrete syntax trees of each source file we extract syntactic hierarchies and integrate them into context window by selectively removing from view more specific, less relevant scopes for a given task. We evaluate this approach on code generation tasks and joint translation of natural language and source code in Python programming language, achieving a new state-of-the-art in code completion and summarization for Python in the CodeXGLUE benchmark. We also introduce new CodeXGLUE benchmarks for user-experience-motivated tasks: code completion with normalized literals, method body completion/code summarization conditioned on file-level context.

Can Language Models be Biomedical Knowledge Bases?
Mujeen Sung et al.

Pre-trained language models (LMs) have become ubiquitous in solving various natural language processing (NLP) tasks. There has been increasing interest in what knowledge these LMs contain and how we can extract that knowledge, treating LMs as knowledge bases (KBs). While there has been much work on probing LMs in the general domain, there has been little attention to whether these powerful LMs can be used as domain-specific KBs. To this end, we create the BioLAMA benchmark, which is comprised of 49K biomedical factual knowledge triples for probing biomedical LMs. We find that biomedical LMs with recently proposed probing methods can achieve up to 18.51% Acc5 on retrieving biomedical knowledge. Although this seems promising given the task difficulty, our detailed analyses reveal that most predictions are highly correlated with prompt templates without any subjects, hence producing similar results on each relation and hindering their capabilities to be used as domain-specific KBs. We hope that BioLAMA can serve as a challenging benchmark for biomedical factual probing.

LayoutReader: Pre-training of Text and Layout for Reading Order Detection
Zilong Wang et al.

Reading order detection is the cornerstone to understanding visually-rich documents (e.g., receipts and forms). Unfortunately, no existing work took advantage of advanced deep learning models because it is too laborious to annotate a large enough dataset. We observe that the reading order of WORD documents is embedded in their XML metadata; meanwhile, it is easy to convert WORD documents to PDFs or images. Therefore, in an automated manner, we construct ReadingBank, a benchmark dataset that contains reading order, text, and layout information for 500,000 document images covering a wide spectrum of document types. This first-ever large-scale dataset unleashes the power of deep neural networks for reading order detection. Specifically, our proposed LayoutReader captures the text and layout information for reading order prediction using the seq2seq model. It performs almost perfectly in reading order detection and significantly improves both open-source and commercial OCR engines in ordering text lines in their results in our experiments. The dataset and models are publicly available at https://aka.ms/layoutreader.

Region under Discussion for visual dialog
Mauricio Mazuecos et al.

Visual Dialog is assumed to require the dialog history to generate correct responses during a dialog. However, it is not clear from previous work how dialog history is needed for visual dialog. In this paper we define what it means for a visual question to require dialog history and we release a subset of the Guesswhat?! questions for which their dialog history completely changes their responses. We propose a novel interpretable representation that visually grounds dialog history: the Region under Discussion. It constrains the image's spatial features according to a semantic representation of the history inspired by the information structure notion of Question under Discussion.We evaluate the architecture on task-specific multimodal models and the visual transformer model LXMERT.

Learning grounded word meaning representations on similarity graphs
Mariella Dimiccoli, Herwig Wendt, and Pau Batlle Franch

This paper introduces a novel approach to learn visually grounded meaning representations of words as low-dimensional node embeddings on an underlying graph hierarchy. The lower level of the hierarchy models modality-specific word representations, conditioned to another modality, through dedicated but communicating graphs, while the higher level puts these representations together on a single graph to learn a representation jointly from both modalities. The topology of each graph models similarity relations among words, and is estimated jointly with the graph embedding. The assumption underlying this model is that words sharing similar meaning correspond to communities in an underlying graph in a low-dimensional space. We named this model Hierarchical Multi-Modal Similarity Graph Embedding (HM-SGE). Experimental results validate the ability of HM-SGE to simulate human similarity judgments and concept categorization, outperforming the state of the art.

WhyAct: Identifying Action Reasons in Lifestyle Vlogs
Oana Ignat et al.

We aim to automatically identify human action reasons in online videos. We focus on the widespread genre of lifestyle vlogs, in which people perform actions while verbally describing them. We introduce and make publicly available the WhyAct dataset, consisting of 1,077 visual actions manually annotated with their reasons. We describe a multimodal model that leverages visual and textual information to automatically infer the reasons corresponding to an action presented in the video.

Genre as Weak Supervision for Cross-lingual Dependency Parsing
Max Müller-Eberstein, Rob van der Goot, and Barbara Plank

Recent work has shown that monolingual masked language models learn to represent data-driven notions of language variation which can be used for domain-targeted training data selection. Dataset genre labels are already frequently available, yet remain largely unexplored in cross-lingual setups. We harness this genre metadata as a weak supervision signal for targeted data selection in zero-shot dependency parsing. Specifically, we project treebank-level genre information to the finer-grained sentence level, with the goal to amplify information implicitly stored in unsupervised contextualized representations. We demonstrate that genre is

recoverable from multilingual contextual embeddings and that it provides an effective signal for training data selection in cross-lingual, zero-shot scenarios. For 12 low-resource language treebanks, six of which are test-only, our genre-specific methods significantly outperform competitive baselines as well as recent embedding-based methods for data selection. Moreover, genre-based data selection provides new state-of-the-art results for three of these target languages.

On the Relation between Syntactic Divergence and Zero-Shot Performance
Ofir Arviv et al.

We explore the link between the extent to which syntactic relations are preserved in translation and the ease of correctly constructing a parse tree in a zero-shot setting. While previous work suggests such a relation, it tends to focus on the macro level and not on the level of individual edges—a gap we aim to address. As a test case, we take the transfer of Universal Dependencies (UD) parsing from English to a diverse set of languages and conduct two sets of experiments. In one, we analyze zero-shot performance based on the extent to which English source edges are preserved in translation. In another, we apply three linguistically motivated transformations to UD, creating more cross-lingually stable versions of it, and assess their zero-shot parsability. In order to compare parsing performance across different schemes, we perform extrinsic evaluation on the downstream task of cross-lingual relation extraction (RE) using a subset of a standard English RE benchmark translated to Russian and Korean. In both sets of experiments, our results suggest a strong relation between cross-lingual stability and zero-shot parsing performance.

Improved Latent Tree Induction with Distant Supervision via Span Constraints
Zhiyang Xu et al.

For over thirty years, researchers have developed and analyzed methods for latent tree induction as an approach for unsupervised syntactic parsing. Nonetheless, modern systems still do not perform well enough compared to their supervised counterparts to have any practical use as structural annotation of text. In this work, we present a technique that uses distant supervision in the form of span constraints (i.e. phrase bracketing) to improve performance in unsupervised constituency parsing. Using a relatively small number of span constraints we can substantially improve the output from DIORA, an already competitive unsupervised parsing system. Compared with full parse tree annotation, span constraints can be acquired with minimal effort, such as with a lexicon derived from Wikipedia, to find exact text matches. Our experiments show span constraints based on entities improves constituency parsing on English WSJ Penn Treebank by more than 5 F1. Furthermore, our method extends to any domain where span constraints are easily attainable, and as a case study we demonstrate its effectiveness by parsing biomedical text from the CRAFT dataset.

Uncovering Main Causalities for Long-tailed Information Extraction
Guoshun Nan et al.

Information Extraction (IE) aims to extract structural information from unstructured texts. In practice, long-tailed distributions caused by the selection bias of a dataset may lead to incorrect correlations, also known as spurious correlations, between entities and labels in the conventional likelihood models. This motivates us to propose counterfactual IE (CFIE), a novel framework that aims to uncover the main causalities behind data in the view of causal inference. Specifically, 1) we first introduce a unified structural causal model (SCM) for various IE tasks, describing the relationships among variables; 2) with our SCM, we then generate counterfactuals based on an explicit language structure to better calculate the direct causal effect during the inference stage; 3) we further propose a novel debiasing approach to yield more robust predictions. Experiments on three IE tasks across five public datasets show the effectiveness of our CFIE model in mitigating the spurious correlation issues.

Separating Retention from Extraction in the Evaluation of End-to-end Relation Extraction
Bruno Taillé et al.

State-of-the-art NLP models can adopt shallow heuristics that limit their generalization capability (McCoy et al., 2019). Such heuristics include lexical overlap with the training set in Named-Entity Recognition (Taille et al., 2020) and Event or Type heuristics in Relation Extraction (Rosenman et al., 2020). In the more realistic end-to-end RE setting, we can expect yet another heuristic: the mere retention of training relation triples. In this paper we propose two experiments confirming that retention of known facts is a key factor of performance on standard benchmarks. Furthermore, one experiment suggests that a pipeline model able to use intermediate type representations is less prone to over-rely on retention.

A Role-Selected Sharing Network for Joint Machine-Human Chatting Handoff and Service Satisfaction Analysis
Jiawei Liu et al.

Chatbot is increasingly thriving in different domains, however, because of unexpected discourse complexity and training data sparseness, its potential distrust hatches vital apprehension. Recently, Machine-Human Chatting Handoff (MHCH), predicting chatbot failure and enabling human-algorithm collaboration to enhance chatbot quality, has attracted increasing attention from industry and academia. In this study, we propose a novel model, Role-Selected Sharing Network (RSSN), which integrates both dialogue satisfaction estimation and handoff prediction in one multi-task learning framework. Unlike prior efforts in dialog mining, by utilizing local user satisfaction as a bridge, global satisfaction detector and handoff predictor can effectively exchange critical information. Specifically, we decouple the relation and interaction between the two tasks by the role information after the shared encoder. Extensive experiments on two public datasets demonstrate the

effectiveness of our model.

Leveraging Order-Free Tag Relations for Context-Aware Recommendation
Junmo Kang et al.

Tag recommendation relies on either a ranking function for top-k tags or an autoregressive generation method. However, the previous methods neglect one of two seemingly conflicting yet desirable characteristics of a tag set: orderlessness and inter-dependency. While the ranking approach fails to address the inter-dependency among tags when they are ranked, the autoregressive approach fails to take orderlessness into account because it is designed to utilize sequential relations among tokens. We propose a sequence-oblivious generation method for tag recommendation, in which the next tag to be generated is independent of the order of the generated tags and the order of the ground truth tags occurring in training data. Empirical results on two different domains, Instagram and Stack Overflow, show that our method is significantly superior to the previous approaches.

Inflate and Shrink:Enriching and Reducing Interactions for Fast Text-Image Retrieval
Haoliang Liu, Tan Yu, and Ping Li

By exploiting the cross-modal attention, cross-BERT methods have achieved state-of-the-art accuracy in cross-modal retrieval. Nevertheless, the heavy text-image interactions in the cross-BERT model are prohibitively slow for large-scale retrieval. Late-interaction methods trade off retrieval accuracy and efficiency by exploiting cross-modal interaction only in the late stage, attaining a satisfactory retrieval speed. In this work, we propose an inflating and shrinking approach to further boost the efficiency and accuracy of late-interaction methods. The inflating operation plugs several codes in the input of the encoder to exploit the text-image interactions more thoroughly for higher retrieval accuracy. Then the shrinking operation gradually reduces the text-image interactions through knowledge distilling for higher efficiency. Through an inflating operation followed by a shrinking operation, both efficiency and accuracy of a late-interaction model are boosted. Systematic experiments on public benchmarks demonstrate the effectiveness of our inflating and shrinking approach.

Vision-and-Language or Vision-for-Language? On Cross-Modal Influence in Multimodal Transformers
Stella Frank, Emanuele Bugliarello, and Desmond Elliott

Pretrained vision-and-language BERTs aim to learn representations that combine information from both modalities. We propose a diagnostic method based on cross-modal input ablation to assess the extent to which these models actually integrate cross-modal information. This method involves ablating inputs from one modality, either entirely or selectively based on cross-modal grounding alignments, and evaluating the model prediction performance on the other modality. Model performance is measured by modality-specific tasks that mirror the model pretraining objectives (e.g. masked language modelling for text). Models that have learned to construct cross-modal representations using both modalities are expected to perform worse when inputs are missing from a modality. We find that recently proposed models have much greater relative difficulty predicting text when visual information is ablated, compared to predicting visual object categories when text is ablated, indicating that these models are not symmetrically cross-modal.

HypMix: Hyperbolic Interpolative Data Augmentation
Ramit Sawhney et al.

Interpolation-based regularisation methods for data augmentation have proven to be effective for various tasks and modalities. These methods involve performing mathematical operations over the raw input samples or their latent states representations - vectors that often possess complex hierarchical geometries. However, these operations are performed in the Euclidean space, simplifying these representations, which may lead to distorted and noisy interpolations. We propose HypMix, a novel model-, data-, and modality-agnostic interpolative data augmentation technique operating in the hyperbolic space, which captures the complex geometry of input and hidden state hierarchies better than its contemporaries. We evaluate HypMix on benchmark and low resource datasets across speech, text, and vision modalities, showing that HypMix consistently outperforms state-of-the-art data augmentation techniques. In addition, we demonstrate the use of HypMix in semi-supervised settings. We further probe into the adversarial robustness and qualitative inferences we draw from HypMix that elucidate the efficacy of the Riemannian hyperbolic manifolds for interpolation-based data augmentation.

Joint Universal Syntactic and Semantic Parsing
Elias Stengel-Eskin et al.

While numerous attempts have been made to jointly parse syntax and semantics, high performance in one domain typically comes at the price of performance in the other. This trade-off contradicts the large body of research focusing on the rich interactions at the syntax-semantics interface. We explore multiple model architectures which allow us to exploit the rich syntactic and semantic annotations contained in the Universal Decompositional Semantics (UDS) dataset, jointly parsing Universal Dependencies and UDS to obtain state-of-the-art results in both formalisms. We analyze the behaviour of a joint model of syntax and semantics, finding patterns supported by linguistic theory at the syntax-semantics interface. We then investigate to what degree joint modeling generalizes to a multilingual setting, where we find similar trends across 8 languages.

In-person Demo Session
Bavaro 4 *09:00–10:30*

Beyond Accuracy: A Consolidated Tool for Visual Question Answering Benchmarking
Dirk Väth, Pascal Tilli, and Ngoc Thang Vu

On the way towards general Visual Question Answering (VQA) systems that are able to answer arbitrary questions, the need arises for evaluation beyond single-metric leaderboards for specific datasets. To this end, we propose a browser-based benchmarking tool for researchers and challenge organizers, with an API for easy integration of new models and datasets to keep up with the fast-changing landscape of VQA. Our tool helps test generalization capabilities of models across multiple datasets, evaluating not just accuracy, but also performance in more realistic real-world scenarios such as robustness to input noise. Additionally, we include metrics that measure biases and uncertainty, to further explain model behavior. Interactive filtering facilitates discovery of problematic behavior, down to the data sample level. As proof of concept, we perform a case study on four models. We find that state-of-the-art VQA models are optimized for specific tasks or datasets, but fail to generalize even to other in-domain test sets, for example they can not recognize text in images. Our metrics allow us to quantify which image and question embeddings provide most robustness to a model. All code s publicly available.

DomiKnowS: A Library for Integration of Symbolic Domain Knowledge in Deep Learning
Hossein Rajaby Faghihi et al.

We demonstrate a library for the integration of domain knowledge in deep learning architectures. Using this library, the structure of the data is expressed symbolically via graph declarations and the logical constraints over outputs or latent variables can be seamlessly added to the deep models. The domain knowledge can be defined explicitly, which improves the explainability of the models in addition to their performance and generalizability in the low-data regime. Several approaches for such integration of symbolic and sub-symbolic models have been introduced; however, there is no library to facilitate the programming for such integration in a generic way while various underlying algorithms can be used. Our library aims to simplify programming for such integration in both training and inference phases while separating the knowledge representation from learning algorithms. We showcase various NLP benchmark tasks and beyond. The framework is publicly available at Github (https://github.com/HLR/DomiKnowS).

Session 6 Overview – Monday, November 8, 2021

Track A	Track B	Track C	Track D	Track E	Track F	
Machine Learning for NLP 4	*Summarization 2*	*Machine Translation and Multilinguality 2*	*Speech, Vision, Robotics, Multimodal Grounding 2*	*Sentiment Analysis, Stylistic Analysis, and Argument Mining 3*	*Semantics 2*	
Editing Factual Knowledge in Language Models *De Cao, Aziz, and Titov*	Aspect-Controllable Opinion Summarization *Amplayo, Angelidis, and Lapata*	Multilingual Unsupervised Neural Machine Translation with Denoising Adapters *Üstün et al.*	Interactive Machine Comprehension with Dynamic Knowledge Graphs *Yuan*	Powering Comparative Classification with Sentiment Analysis via Domain Adaptive Knowledge Transfer *Li et al.*	SimCSE: Simple Contrastive Learning of Sentence Embeddings *Gao, Yao, and Chen*	2:45
Sparse Attention with Linear Units *Zhang, Titov, and Sennrich*	QuestEval: Summarization Asks for Fact-based Evaluation *Scialom et al.*	BERT, mBERT, or BiBERT? A Study on Contextualized Embeddings for Neural Machine Translation *Xu, Van Durme, and Murray*	Residual Adapters for Parameter-Efficient ASR Adaptation to Atypical and Accented Speech *Tomanek et al.*	Tribrid: Stance Classification with Neural Inconsistency Detection *Yang and Urbani*	When is Wall a Pared and when a Muro?: Extracting Rules Governing Lexical Selection *Chaudhary et al.*	3:00
Knowledge Base Completion Meets Transfer Learning *Kocijan and Lukasiewicz*	Simple Conversational Data Augmentation for Semi-supervised Abstractive Dialogue Summarization *Chen and Yang*	Controlling Machine Translation for Multiple Attributes with Additive Interventions *Schioppa et al.*	Visual News: Benchmark and Challenges in News Image Captioning *Liu et al.*	SYSML: StYlometry with Structure and Multi-task Learning: Implications for Darknet Forum Migrant Analysis *Maneriker, He, and Parthasarathy*	Aligning Actions Across Recipe Graphs *Donatelli et al.*	3:15
SPECTRA: Sparse Structured Text Rationalization *Guerreiro and Martins*	Finding a Balanced Degree of Automation for Summary Evaluation *Zhang and Bansal*	A Generative Framework for Simultaneous Machine Translation *Miao, Blunsom, and Specia*	Integrating Visuospatial, Linguistic, and Commonsense Structure into Story Visualization *Maharana and Bansal*	Few-Shot Emotion Recognition in Conversation with Sequential Prototypical Networks *Guibon et al.*	Iterative Paraphrastic Augmentation with Discriminative Span Alignment *Van Durme et al.*	3:30

	Track A	**Track B**	**Track C**	**Track D**	**Track E**	**Track F**
	Machine Learning for NLP 4	*Summarization 2*	*Machine Translation and Multilinguality 2*	*Speech, Vision, Robotics, Multimodal Grounding 2*	*Sentiment Analysis, Stylistic Analysis, and Argument Mining 3*	*Semantics 2*
3:45	Towards Zero-Shot Knowledge Distillation for Natural Language Processing *Rashid et al.*	CLIFF: Contrastive Learning for Improving Faithfulness and Factuality in Abstractive Summarization *Cao and Wang*	It Is Not As Good As You Think! Evaluating Simultaneous Machine Translation on Interpretation Data *Zhao et al.*	VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding *Xu et al.*	CLASSIC: Continual and Contrastive Learning of Aspect Sentiment Classification Tasks *Ke et al.*	Generating Datasets with Pretrained Language Models *Schick and Schütze*
4:00	Adversarial Regularization as Stackelberg Game: An Unrolled Optimization Approach *Zuo et al.*	A Statistical Analysis of Summarization Evaluation Metrics Using Resampling Methods *Deutsch, Dror, and Roth*	Boosting Cross-Lingual Transfer via Self-Learning with Uncertainty Estimation *Xu et al.*	NewsCLIPpings: Automatic Generation of Out-of-Context Multimodal Media *Luo, Darrell, and Rohrbach*	Implicit Sentiment Analysis with Event-centered Text Representation *Zhou et al.*	Continuous Entailment Patterns for Lexical Inference in Context *Schmitt and Schütze*
4:05			Levenshtein Training for Word-level Quality Estimation *Ding et al.*			Numeracy enhances the Literacy of Language Models *Thawani, Pujara, and Ilievski*

Parallel Session 6

Session 6A: Machine Learning for NLP 4

Bavaro 1 *Chair: Karthik Narasimhan*

Editing Factual Knowledge in Language Models

Nicola De Cao, Wilker Aziz, and Ivan Titov 14:45–15:00

The factual knowledge acquired during pre-training and stored in the parameters of Language Models (LMs) can be useful in downstream tasks (e.g., question answering or textual inference). However, some facts can be incorrectly induced or become obsolete over time. We present KnowledgeEditor, a method which can be used to edit this knowledge and, thus, fix 'bugs' or unexpected predictions without the need for expensive re-training or fine-tuning. Besides being computationally efficient, KnowledgeEditordoes not require any modifications in LM pre-training (e.g., the use of meta-learning). In our approach, we train a hyper-network with constrained optimization to modify a fact without affecting the rest of the knowledge; the trained hyper-network is then used to predict the weight update at test time. We show KnowledgeEditor's efficacy with two popular architectures and knowledge-intensive tasks: i) a BERT model fine-tuned for fact-checking, and ii) a sequence-to-sequence BART model for question answering. With our method, changing a prediction on the specific wording of a query tends to result in a consistent change in predictions also for its paraphrases. We show that this can be further encouraged by exploiting (e.g., automatically-generated) paraphrases during training. Interestingly, our hyper-network can be regarded as a 'probe' revealing which components need to be changed to manipulate factual knowledge; our analysis shows that the updates tend to be concentrated on a small subset of components. Source code available at https://github.com/nicola-decao/KnowledgeEditor

Sparse Attention with Linear Units

Biao Zhang, Ivan Titov, and Rico Sennrich 15:00–15:15

Recently, it has been argued that encoder-decoder models can be made more interpretable by replacing the softmax function in the attention with its sparse variants. In this work, we introduce a novel, simple method for achieving sparsity in attention: we replace the softmax activation with a , and show that sparsity naturally emerges from such a formulation. Training stability is achieved with layer normalization with either a specialized initialization or an additional gating function. Our model, which we call Rectified Linear Attention (ReLA), is easy to implement and more efficient than previously proposed sparse attention mechanisms. We apply ReLA to the Transformer and conduct experiments on five machine translation tasks. ReLA achieves translation performance comparable to several strong baselines, with training and decoding speed similar to that of the vanilla attention. Our analysis shows that ReLA delivers high sparsity rate and head diversity, and the induced cross attention achieves better accuracy with respect to source-target word alignment than recent sparsified softmax-based models. Intriguingly, ReLA heads also learn to attend to nothing (i.e. 'switch off') for some queries, which is not possible with sparsified softmax alternatives.

Knowledge Base Completion Meets Transfer Learning

Vid Kocijan and Thomas Lukasiewicz 15:15–15:30

The aim of knowledge base completion is to predict unseen facts from existing facts in knowledge bases. In this work, we introduce the first approach for transfer of knowledge from one collection of facts to another without the need for entity or relation matching. The method works for both canonicalized knowledge bases and uncanonicalized or open knowledge bases, i.e., knowledge bases where more than one copy of a real-world entity or relation may exist. Such knowledge bases are a natural output of automated information extraction tools that extract structured data from unstructured text. Our main contribution is a method that can make use of a large-scale pretraining on facts, collected from unstructured text, to improve predictions on structured data from a specific domain. The introduced method is the most impactful on small datasets such as ReVerb20K, where we obtained a 6% absolute increase of mean reciprocal rank and 65% relative decrease of mean rank over the previously best method, despite not relying on large pre-trained models like BERT.

SPECTRA: Sparse Structured Text Rationalization

Nuno M. Guerreiro and André F. T. Martins 15:30–15:45

Selective rationalization aims to produce decisions along with rationales (e.g., text highlights or word alignments between two sentences). Commonly, rationales are modeled as stochastic binary masks, requiring sampling-based gradient estimators, which complicates training and requires careful hyperparameter tuning. Sparse attention mechanisms are a deterministic alternative, but they lack a way to regularize the rationale extraction (e.g., to control the sparsity of a text highlight or the number of alignments). In this paper, we present a unified framework for deterministic extraction of structured explanations via constrained inference on a factor graph, forming a differentiable layer. Our approach greatly eases training and rationale regularization, generally outperforming previous work on what comes to performance and plausibility of the extracted rationales. We further provide a comparative study of stochastic and deterministic methods for rationale extraction for classification and natural language inference tasks, jointly assessing their predictive power, quality of the explanations, and model variability.

Towards Zero-Shot Knowledge Distillation for Natural Language Processing

Ahmad Rashid et al. 15:45–16:00

Knowledge distillation (KD) is a common knowledge transfer algorithm used for model compression across a variety of deep learning based natural language processing (NLP) solutions. In its regular manifestations, KD requires access to the teacher's training data for knowledge transfer to the student network. However, privacy concerns, data regulations and proprietary reasons may prevent access to such data. We present, to the best of our knowledge, the first work on Zero-shot Knowledge Distillation for NLP, where the student learns from the much larger teacher without any task specific data. Our solution combines out-of-domain data and adversarial training to learn the teacher's output distribution. We investigate six tasks from the GLUE benchmark and demonstrate that we can achieve between 75% and 92% of the teacher's classification score (accuracy or F1) while compressing the model 30 times.

Adversarial Regularization as Stackelberg Game: An Unrolled Optimization Approach
Simiao Zuo et al. 16:00–16:15

Adversarial regularization has been shown to improve the generalization performance of deep learning models in various natural language processing tasks. Existing works usually formulate the method as a zero-sum game, which is solved by alternating gradient descent/ascent algorithms. Such a formulation treats the adversarial and the defending players equally, which is undesirable because only the defending player contributes to the generalization performance. To address this issue, we propose Stackelberg Adversarial Regularization (SALT), which formulates adversarial regularization as a Stackelberg game. This formulation induces a competition between a leader and a follower, where the follower generates perturbations, and the leader trains the model subject to the perturbations. Different from conventional approaches, in SALT, the leader is in an advantageous position. When the leader moves, it recognizes the strategy of the follower and takes the anticipated follower's outcomes into consideration. Such a leader's advantage enables us to improve the model fitting to the unperturbed data. The leader's strategic information is captured by the Stackelberg gradient, which is obtained using an unrolling algorithm. Our experimental results on a set of machine translation and natural language understanding tasks show that SALT outperforms existing adversarial regularization baselines across all tasks. Our code is publicly available.

Session 6B: Summarization 2

Bavaro 2 *Chair: Rui Zhang*

Aspect-Controllable Opinion Summarization

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata 14:45–15:00

Recent work on opinion summarization produces general summaries based on a set of input reviews and the popularity of opinions expressed in them. In this paper, we propose an approach that allows the generation of customized summaries based on aspect queries (e.g., describing the location and room of a hotel). Using a review corpus, we create a synthetic training dataset of (review, summary) pairs enriched with aspect controllers which are induced by a multi-instance learning model that predicts the aspects of a document at different levels of granularity. We fine-tune a pretrained model using our synthetic dataset and generate aspect-specific summaries by modifying the aspect controllers. Experiments on two benchmarks show that our model outperforms the previous state of the art and generates personalized summaries by controlling the number of aspects discussed in them.

QuestEval: Summarization Asks for Fact-based Evaluation

Thomas Scialom et al. 15:00–15:15

Summarization evaluation remains an open research problem: current metrics such as ROUGE are known to be limited and to correlate poorly with human judgments. To alleviate this issue, recent work has proposed evaluation metrics which rely on question answering models to assess whether a summary contains all the relevant information in its source document. Though promising, the proposed approaches have so far failed to correlate better than ROUGE with human judgments. In this paper, we extend previous approaches and propose a unified framework, named QuestEval. In contrast to established metrics such as ROUGE or BERTScore, QuestEval does not require any ground-truth reference. Nonetheless, QuestEval substantially improves the correlation with human judgments over four evaluation dimensions (consistency, coherence, fluency, and relevance), as shown in extensive experiments.

Simple Conversational Data Augmentation for Semi-supervised Abstractive Dialogue Summarization

Jiaao Chen and Diyi Yang 15:15–15:30

Abstractive conversation summarization has received growing attention while most current state-of-the-art summarization models heavily rely on human-annotated summaries. To reduce the dependence on labeled summaries, in this work, we present a simple yet effective set of Conversational Data Augmentation (CODA) methods for semi-supervised abstractive conversation summarization, such as random swapping/deletion to perturb the discourse relations inside conversations, dialogue-acts-guided insertion to interrupt the development of conversations, and conditional-generation-based substitution to substitute utterances with their paraphrases generated based on the conversation context. To further utilize unlabeled conversations, we combine CODA with two-stage noisy self-training where we first pre-train the summarization model on unlabeled conversations with pseudo summaries and then fine-tune it on labeled conversations. Experiments conducted on the recent conversation summarization datasets demonstrate the effectiveness of our methods over several state-of-the-art data augmentation baselines.

Finding a Balanced Degree of Automation for Summary Evaluation

Shiyue Zhang and Mohit Bansal 15:30–15:45

Human evaluation for summarization tasks is reliable but brings in issues of reproducibility and high costs. Automatic metrics are cheap and reproducible but sometimes poorly correlated with human judgment. In this work, we propose flexible semiautomatic to automatic summary evaluation metrics, following the Pyramid human evaluation method. Semi-automatic Lite2Pyramid retains the reusable human-labeled Summary Content Units (SCUs) for reference(s) but replaces the manual work of judging SCUs' presence in system summaries with a natural language inference (NLI) model. Fully automatic Lite3Pyramid further substitutes SCUs with automatically extracted Semantic Triplet Units (STUs) via a semantic role labeling (SRL) model. Finally, we propose in-between metrics, Lite2.xPyramid, where we use a simple regressor to predict how well the STUs can simulate SCUs and retain SCUs that are more difficult to simulate, which provides a smooth transition and balance between automation and manual evaluation. Comparing to 15 existing metrics, we evaluate human-metric correlations on 3 existing meta-evaluation datasets and our newly collected PyrXSum (with 100/10 XSum examples/systems). It shows that Lite2Pyramid consistently has the best summary-level correlations; Lite3Pyramid works better than or comparable to other automatic metrics; Lite2.xPyramid trades off small correlation drops for larger manual effort reduction, which can reduce costs for future data collection.

CLIFF: Contrastive Learning for Improving Faithfulness and Factuality in Abstractive Summarization

Shuyang Cao and Lu Wang 15:45–16:00

We study generating abstractive summaries that are faithful and factually consistent with the given articles. A novel contrastive learning formulation is presented, which leverages both reference summaries, as positive training data, and automatically generated erroneous summaries, as negative training data, to train summarization systems that are better at distinguishing between them. We further design four types of strategies for creating negative samples, to resemble errors made commonly by two state-of-the-art models, BART and PEGASUS, found in our new human annotations of summary errors. Experiments on XSum and CNN/Daily Mail

show that our contrastive learning framework is robust across datasets and models. It consistently produces more factual summaries than strong comparisons with post error correction, entailment-based reranking, and unlikelihood training, according to QA-based factuality evaluation. Human judges echo the observation and find that our model summaries correct more errors.

A Statistical Analysis of Summarization Evaluation Metrics Using Resampling Methods
Daniel Deutsch, Rotem Dror, and Dan Roth 16:00–16:15

The quality of a summarization evaluation metric is quantified by calculating the correlation between its scores and human annotations across a large number of summaries. Currently, it is unclear how precise these correlation estimates are, nor whether differences between two metrics' correlations reflect a true difference or if it is due to mere chance. In this work, we address these two problems by proposing methods for calculating confidence intervals and running hypothesis tests for correlations using two resampling methods, bootstrapping and permutation. After evaluating which of the proposed methods is most appropriate for summarization through two simulation experiments, we analyze the results of applying these methods to several different automatic evaluation metrics across three sets of human annotations. We find that the confidence intervals are rather wide, demonstrating high uncertainty in the reliability of automatic metrics. Further, although many metrics fail to show statistical improvements over ROUGE, two recent works, QAEval and BERTScore, do in some evaluation settings.

Session 6C: Machine Translation and Multilinguality 2

Bavaro 3 *Chair: Rico Sennrich*

Multilingual Unsupervised Neural Machine Translation with Denoising Adapters

Ahmet Üstün et al. 14:45–15:00

We consider the problem of multilingual unsupervised machine translation, translating to and from languages that only have monolingual data by using auxiliary parallel language pairs. For this problem the standard procedure so far to leverage the monolingual data is _back-translation_, which is computationally costly and hard to tune. In this paper we propose instead to use _denoising adapters_, adapter layers with a denoising objective, on top of pre-trained mBART-50. In addition to the modularity and flexibility of such an approach we show that the resulting translations are on-par with back-translating as measured by BLEU, and furthermore it allows adding unseen languages incrementally.

BERT, mBERT, or BiBERT? A Study on Contextualized Embeddings for Neural Machine Translation

Haoran Xu, Benjamin Van Durme, and Kenton Murray 15:00–15:15

The success of bidirectional encoders using masked language models, such as BERT, on numerous natural language processing tasks has prompted researchers to attempt to incorporate these pre-trained models into neural machine translation (NMT) systems. However, proposed methods for incorporating pre-trained models are non-trivial and mainly focus on BERT, which lacks a comparison of the impact that other pre-trained models may have on translation performance. In this paper, we demonstrate that simply using the output (contextualized embeddings) of a tailored and suitable bilingual pre-trained language model (dubbed BiBERT) as the input of the NMT encoder achieves state-of-the-art translation performance. Moreover, we also propose a stochastic layer selection approach and a concept of a dual-directional translation model to ensure the sufficient utilization of contextualized embeddings. In the case of without using back translation, our best models achieve BLEU scores of 30.45 for En→De and 38.61 for De→En on the IWSLT'14 dataset, and 31.26 for En→De and 34.94 for De→En on the WMT'14 dataset, which exceeds all published numbers.

Controlling Machine Translation for Multiple Attributes with Additive Interventions

Andrea Schioppa et al. 15:15–15:30

Fine-grained control of machine translation (MT) outputs along multiple attributes is critical for many modern MT applications and is a requirement for gaining users' trust. A standard approach for exerting control in MT is to prepend the input with a special tag to signal the desired output attribute. Despite its simplicity, attribute tagging has several drawbacks: continuous values must be binned into discrete categories, which is unnatural for certain applications; interference between multiple tags is poorly understood. We address these problems by introducing vector-valued interventions which allow for fine-grained control over multiple attributes simultaneously via a weighted linear combination of the corresponding vectors. For some attributes, our approach even allows for fine-tuning a model trained without annotations to support such interventions. In experiments with three attributes (length, politeness and monotonicity) and two language pairs (English to German and Japanese) our models achieve better control over a wider range of tasks compared to tagging, and translation quality does not degrade when no control is requested. Finally, we demonstrate how to enable control in an already trained model after a relatively cheap fine-tuning stage.

A Generative Framework for Simultaneous Machine Translation

Yishu Miao, Phil Blunsom, and Lucia Specia 15:30–15:45

We propose a generative framework for simultaneous machine translation. Conventional approaches use a fixed number of source words to translate or learn dynamic policies for the number of source words by reinforcement learning. Here we formulate simultaneous translation as a structural sequence-to-sequence learning problem. A latent variable is introduced to model read or translate actions at every time step, which is then integrated out to consider all the possible translation policies. A re-parameterised Poisson prior is used to regularise the policies which allows the model to explicitly balance translation quality and latency. The experiments demonstrate the effectiveness and robustness of the generative framework, which achieves the best BLEU scores given different average translation latencies on benchmark datasets.

It Is Not As Good As You Think! Evaluating Simultaneous Machine Translation on Interpretation Data

Jinming Zhao et al. 15:45–15:55

Most existing simultaneous machine translation (SiMT) systems are trained and evaluated on offline translation corpora. We argue that SiMT systems should be trained and tested on real interpretation data. To illustrate this argument, we propose an interpretation test set and conduct a realistic evaluation of SiMT trained on offline translations. Our results, on our test set along with 3 existing smaller scale language pairs, highlight the difference of up-to 13.83 BLEU score when SiMT models are evaluated on translation vs interpretation data. In the absence of interpretation training data, we propose a translation-to-interpretation (T2I) style transfer method which allows converting existing offline translations into interpretation-style data, leading to up-to 2.8 BLEU improvement. However, the evaluation gap remains notable, calling for constructing large-scale interpretation corpora better suited for evaluating and developing SiMT systems.

Boosting Cross-Lingual Transfer via Self-Learning with Uncertainty Estimation
Liyan Xu et al. 15:55–16:05

Recent multilingual pre-trained language models have achieved remarkable zero-shot performance, where the model is only finetuned on one source language and directly evaluated on target languages. In this work, we propose a self-learning framework that further utilizes unlabeled data of target languages, combined with uncertainty estimation in the process to select high-quality silver labels. Three different uncertainties are adapted and analyzed specifically for the cross lingual transfer: Language Heteroscedastic/Homoscedastic Uncertainty (LEU/LOU), Evidential Uncertainty (EVI). We evaluate our framework with uncertainties on two cross-lingual tasks including Named Entity Recognition (NER) and Natural Language Inference (NLI) covering 40 languages in total, which outperforms the baselines significantly by 10 F1 for NER on average and 2.5 accuracy for NLI.

Levenshtein Training for Word-level Quality Estimation
Shuoyang Ding et al. 16:05–16:15

We propose a novel scheme to use the Levenshtein Transformer to perform the task of word-level quality estimation. A Levenshtein Transformer is a natural fit for this task: trained to perform decoding in an iterative manner, a Levenshtein Transformer can learn to post-edit without explicit supervision. To further minimize the mismatch between the translation task and the word-level QE task, we propose a two-stage transfer learning procedure on both augmented data and human post-editing data. We also propose heuristics to construct reference labels that are compatible with subword-level finetuning and inference. Results on WMT 2020 QE shared task dataset show that our proposed method has superior data efficiency under the data-constrained setting and competitive performance under the unconstrained setting.

Session 6D: Speech, Vision, Robotics, Multimodal Grounding 2

Punta Cana 1,2 *Chair: Pranava Madhyastha*

Interactive Machine Comprehension with Dynamic Knowledge Graphs

Xingdi Yuan 14:45–15:00

Interactive machine reading comprehension (iMRC) is machine comprehension tasks where knowledge sources are partially observable. An agent must interact with an environment sequentially to gather necessary knowledge in order to answer a question. We hypothesize that graph representations are good inductive biases, which can serve as an agent's memory mechanism in iMRC tasks. We explore four different categories of graphs that can capture text information at various levels. We describe methods that dynamically build and update these graphs during information gathering, as well as neural models to encode graph representations in RL agents. Extensive experiments on iSQuAD suggest that graph representations can result in significant performance improvements for RL agents.

Residual Adapters for Parameter-Efficient ASR Adaptation to Atypical and Accented Speech

Katrin Tomanek et al. 15:00–15:15

Automatic Speech Recognition (ASR) systems are often optimized to work best for speakers with canonical speech patterns. Unfortunately, these systems perform poorly when tested on atypical speech and heavily accented speech. It has previously been shown that personalization through model fine-tuning substantially improves performance. However, maintaining such large models per speaker is costly and difficult to scale. We show that by adding a relatively small number of extra parameters to the encoder layers via so-called residual adapter, we can achieve similar adaptation gains compared to model fine-tuning, while only updating a tiny fraction (less than 0.5%) of the model parameters. We demonstrate this on two speech adaptation tasks (atypical and accented speech) and for two state-of-the-art ASR architectures.

Visual News: Benchmark and Challenges in News Image Captioning

Fuxiao Liu et al. 15:15–15:30

We propose Visual News Captioner, an entity-aware model for the task of news image captioning. We also introduce Visual News, a large-scale benchmark consisting of more than one million news images along with associated news articles, image captions, author information, and other metadata. Unlike the standard image captioning task, news images depict situations where people, locations, and events are of paramount importance. Our proposed method can effectively combine visual and textual features to generate captions with richer information such as events and entities. More specifically, built upon the Transformer architecture, our model is further equipped with novel multi-modal feature fusion techniques and attention mechanisms, which are designed to generate named entities more accurately. Our method utilizes much fewer parameters while achieving slightly better prediction results than competing methods. Our larger and more diverse Visual News dataset further highlights the remaining challenges in captioning news images.

Integrating Visuospatial, Linguistic, and Commonsense Structure into Story Visualization

Adyasha Maharana and Mohit Bansal 15:30–15:45

While much research has been done in text-to-image synthesis, little work has been done to explore the usage of linguistic structure of the input text. Such information is even more important for story visualization since its inputs have an explicit narrative structure that needs to be translated into an image sequence (or visual story). Prior work in this domain has shown that there is ample room for improvement in the generated image sequence in terms of visual quality, consistency and relevance. In this paper, we first explore the use of constituency parse trees using a Transformer-based recurrent architecture for encoding structured input. Second, we augment the structured input with commonsense information and study the impact of this external knowledge on the generation of visual story. Third, we also incorporate visual structure via bounding boxes and dense captioning to provide feedback about the characters/objects in generated images within a dual learning setup. We show that off-the-shelf dense-captioning models trained on Visual Genome can improve the spatial structure of images from a different target domain without needing fine-tuning. We train the model end-to-end using intra-story contrastive loss (between words and image sub-regions) and show significant improvements in visual quality. Finally, we provide an analysis of the linguistic and visuo-spatial information.

VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding

Hu Xu et al. 15:45–16:00

We present VideoCLIP, a contrastive approach to pre-train a unified model for zero-shot video and text understanding, without using any labels on downstream tasks. VideoCLIP trains a transformer for video and text by contrasting temporally overlapping positive video-text pairs with hard negatives from nearest neighbor retrieval. Our experiments on a diverse series of downstream tasks, including sequence-level text-video retrieval, VideoQA, token-level action localization, and action segmentation reveal state-of-the-art performance, surpassing prior work, and in some cases even outperforming supervised approaches. Code is made available at https://github.com/pytorch/fairseq/examples/MMPT.

NewsCLIPpings: Automatic Generation of Out-of-Context Multimodal Media

Grace Luo, Trevor Darrell, and Anna Rohrbach 16:00–16:15

Online misinformation is a prevalent societal issue, with adversaries relying on tools ranging from cheap fakes to sophisticated deep fakes. We are motivated by the threat scenario where an image is used out of context to support a certain narrative. While some prior datasets for detecting image-text inconsistency generate samples via text manipulation, we propose a dataset where both image and text are unmanipulated but mismatched. We introduce several strategies for automatically retrieving convincing images for a given caption, capturing cases with inconsistent entities or semantic context. Our large-scale automatically generated the NewsCLIPpings Dataset: (1) demonstrates that machine-driven image repurposing is now a realistic threat, and (2) provides samples that represent challenging instances of mismatch between text and image in news that are able to mislead humans. We benchmark several state-of-the-art multimodal models on our dataset and analyze their performance across different pretraining domains and visual backbones.

Session 6E: Sentiment Analysis, Stylistic Analysis, and Argument Mining 3

Punta Cana 3,4 *Chair: Zhongyu Wei*

Powering Comparative Classification with Sentiment Analysis via Domain Adaptive Knowledge Transfer

Zeyu Li et al. 16:45–17:00

We study Comparative Preference Classification (CPC) which aims at predicting whether a preference comparison exists between two entities in a given sentence and, if so, which entity is preferred over the other. High-quality CPC models can significantly benefit applications such as comparative question answering and review-based recommendation. Among the existing approaches, non-deep learning methods suffer from inferior performances. The state-of-the-art graph neural network-based ED-GAT (Ma et al., 2020) only considers syntactic information while ignoring the critical semantic relations and the sentiments to the compared entities. We propose Sentiment Analysis Enhanced COmparative Network (SAECON) which improves CPC accuracy with a sentiment analyzer that learns sentiments to individual entities via domain adaptive knowledge transfer. Experiments on the CompSent-19 (Panchenko et al., 2019) dataset present a significant improvement on the F1 scores over the best existing CPC approaches.

Tribrid: Stance Classification with Neural Inconsistency Detection

Song Yang and Jacopo Urbani 17:00–17:15

We study the problem of performing automatic stance classification on social media with neural architectures such as BERT. Although these architectures deliver impressive results, their level is not yet comparable to the one of humans and they might produce errors that have a significant impact on the downstream task (e.g., fact-checking). To improve the performance, we present a new neural architecture where the input also includes automatically generated negated perspectives over a given claim. The model is jointly learned to make simultaneously multiple predictions, which can be used either to improve the classification of the original perspective or to filter out doubtful predictions. In the first case, we propose a weakly supervised method for combining the predictions into a final one. In the second case, we show that using the confidence scores to remove doubtful predictions allows our method to achieve human-like performance over the retained information, which is still a sizable part of the original input.

SYSML: StYlometry with Structure and Multitask Learning: Implications for Darknet Forum Migrant Analysis

Pranav Maneriker, Yuntian He, and Srinivasan Parthasarathy 17:15–17:30

Darknet market forums are frequently used to exchange illegal goods and services between parties who use encryption to conceal their identities. The Tor network is used to host these markets, which guarantees additional anonymization from IP and location tracking, making it challenging to link across malicious users using multiple accounts (sybils). Additionally, users migrate to new forums when one is closed further increasing the difficulty of linking users across multiple forums. We develop a novel stylometry-based multitask learning approach for natural language and model interactions using graph embeddings to construct low-dimensional representations of short episodes of user activity for authorship attribution. We provide a comprehensive evaluation of our methods across four different darknet forums demonstrating its efficacy over the state-of-the-art, with a lift of up to 2.5X on Mean Retrieval Rank and 2X on Recall10.

Few-Shot Emotion Recognition in Conversation with Sequential Prototypical Networks

Gaël Guibon et al. 17:30–17:45

Several recent studies on dyadic human-human interactions have been done on conversations without specific business objectives. However, many companies might benefit from studies dedicated to more precise environments such as after sales services or customer satisfaction surveys. In this work, we place ourselves in the scope of a live chat customer service in which we want to detect emotions and their evolution in the conversation flow. This context leads to multiple challenges that range from exploiting restricted, small and mostly unlabeled datasets to finding and adapting methods for such context. We tackle these challenges by using Few-Shot Learning while making the hypothesis it can serve conversational emotion classification for different languages and sparse labels. We contribute by proposing a variation of Prototypical Networks for sequence labeling in conversation that we name ProtoSeq. We test this method on two datasets with different languages: daily conversations in English and customer service chat conversations in French. When applied to emotion classification in conversations, our method proved to be competitive even when compared to other ones.

CLASSIC: Continual and Contrastive Learning of Aspect Sentiment Classification Tasks

Zixuan Ke et al. 17:45–18:00

This paper studies continual learning (CL) of a sequence of aspect sentiment classification (ASC) tasks in a particular CL setting called domain incremental learning (DIL). Each task is from a different domain or product. The DIL setting is particularly suited to ASC because in testing the system needs not know the task/domain to which the test data belongs. To our knowledge, this setting has not been studied before for ASC. This paper proposes a novel model called CLASSIC. The key novelty is a contrastive continual learning method that enables both knowledge transfer across tasks and knowledge distillation from old tasks to the new task, which eliminates the need for task ids in testing. Experimental results show the high effectiveness of

CLASSIC.

Implicit Sentiment Analysis with Event-centered Text Representation
Deyu Zhou et al. 18:00–18:15

Implicit sentiment analysis, aiming at detecting the sentiment of a sentence without sentiment words, has become an attractive research topic in recent years. In this paper, we focus on event-centric implicit sentiment analysis that utilizes the sentiment-aware event contained in a sentence to infer its sentiment polarity. Most existing methods in implicit sentiment analysis simply view noun phrases or entities in text as events or indirectly model events with sophisticated models. Since events often trigger sentiments in sentences, we argue that this task would benefit from explicit modeling of events and event representation learning. To this end, we represent an event as the combination of its event type and the event triplet <subject, predicate, object>. Based on such event representation, we further propose a novel model with hierarchical tensor-based composition mechanism to detect sentiment in text. In addition, we present a dataset for event-centric implicit sentiment analysis where each sentence is labeled with the event representation described above. Experimental results on our constructed dataset and an existing benchmark dataset show the effectiveness of the proposed approach.

Session 6F: Semantics 2

Santo Domingo *Chair: Diyi Yang*

SimCSE: Simple Contrastive Learning of Sentence Embeddings

Tianyu Gao, Xingcheng Yao, and Danqi Chen 14:45–15:00

This paper presents SimCSE, a simple contrastive learning framework that greatly advances the state-of-the-art sentence embeddings. We first describe an unsupervised approach, which takes an input sentence and predicts itself in a contrastive objective, with only standard dropout used as noise. This simple method works surprisingly well, performing on par with previous supervised counterparts. We find that dropout acts as minimal data augmentation and removing it leads to a representation collapse. Then, we propose a supervised approach, which incorporates annotated pairs from natural language inference datasets into our contrastive learning framework, by using "entailment" pairs as positives and "contradiction" pairs as hard negatives. We evaluate SimCSE on standard semantic textual similarity (STS) tasks, and our unsupervised and supervised models using BERT base achieve an average of 76.3% and 81.6% Spearman's correlation respectively, a 4.2% and 2.2% improvement compared to previous best results. We also show—both theoretically and empirically—that contrastive learning objective regularizes pre-trained embeddings' anisotropic space to be more uniform, and it better aligns positive pairs when supervised signals are available.

When is Wall a Pared and when a Muro?: Extracting Rules Governing Lexical Selection

Aditi Chaudhary et al. 15:00–15:15

Learning fine-grained distinctions between vocabulary items is a key challenge in learning a new language. For example, the noun "wall" has different lexical manifestations in Spanish – "pared" refers to an indoor wall while "muro" refers to an outside wall. However, this variety of lexical distinction may not be obvious to non-native learners unless the distinction is explained in such a way. In this work, we present a method for automatically identifying fine-grained lexical distinctions, and extracting rules explaining these distinctions in a human- and machine-readable format. We confirm the quality of these extracted rules in a language learning setup for two languages, Spanish and Greek, where we use the rules to teach non-native speakers when to translate a given ambiguous word into its different possible translations.

Aligning Actions Across Recipe Graphs

Lucia Donatelli et al. 15:15–15:30

Recipe texts are an idiosyncratic form of instructional language that pose unique challenges for automatic understanding. One challenge is that a cooking step in one recipe can be explained in another recipe in different words, at a different level of abstraction, or not at all. Previous work has annotated correspondences between recipe instructions at the sentence level, often glossing over important correspondences between cooking steps across recipes. We present a novel and fully-parsed English recipe corpus, ARA (Aligned Recipe Actions), which annotates correspondences between individual actions across similar recipes with the goal of capturing information implicit for accurate recipe understanding. We represent this information in the form of recipe graphs, and we train a neural model for predicting correspondences on ARA. We find that substantial gains in accuracy can be obtained by taking fine-grained structural information about the recipes into account.

Iterative Paraphrastic Augmentation with Discriminative Span Alignment

Benjamin Van Durme et al. 15:30–15:45

We introduce a novel paraphrastic augmentation strategy based on sentence-level lexically constrained paraphrasing and discriminative span alignment. Our approach allows for the large-scale expansion of existing datasets or the rapid creation of new datasets using a small, manually-produced seed corpus. We demonstrate our approach with experiments on the Berkeley FrameNet Project, a large-scale language understanding effort spanning more than two decades of human labor. With four days of training data collection for a span alignment model and one day of parallel compute, we automatically generate and release to the community 495,300 unique (Frame, Trigger) pairs in diverse sentential contexts, a roughly 50-fold expansion atop FrameNet v1.7. The resulting dataset is intrinsically and extrinsically evaluated in detail, showing positive results on a downstream task.

Generating Datasets with Pretrained Language Models

Timo Schick and Hinrich Schütze 15:45–15:55

To obtain high-quality sentence embeddings from pretrained language models (PLMs), they must either be augmented with additional pretraining objectives or finetuned on a large set of labeled text pairs. While the latter approach typically outperforms the former, it requires great human effort to generate suitable datasets of sufficient size. In this paper, we show how PLMs can be leveraged to obtain high-quality sentence embeddings without the need for labeled data, finetuning or modifications to the pretraining objective: We utilize the generative abilities of large and high-performing PLMs to generate entire datasets of labeled text pairs from scratch, which we then use for finetuning much smaller and more efficient models. Our fully unsupervised approach outperforms strong baselines on several semantic textual similarity datasets.

Continuous Entailment Patterns for Lexical Inference in Context

Martin Schmitt and Hinrich Schütze 15:55–16:05

Combining a pretrained language model (PLM) with textual patterns has been shown to help in both zero- and few-shot settings. For zero-shot performance, it makes sense to design patterns that closely resemble

the text seen during self-supervised pretraining because the model has never seen anything else. Supervised training allows for more flexibility. If we allow for tokens outside the PLM's vocabulary, patterns can be adapted more flexibly to a PLM's idiosyncrasies. Contrasting patterns where a "token" can be any continuous vector from those where a discrete choice between vocabulary elements has to be made, we call our method CONtinous pAtterNs (CONAN). We evaluate CONAN on two established benchmarks for lexical inference in context (LIiC) a.k.a. predicate entailment, a challenging natural language understanding task with relatively small training data. In a direct comparison with discrete patterns, CONAN consistently leads to improved performance, setting a new state of the art. Our experiments give valuable insights on the kind of pattern that enhances a PLM's performance on LIiC and raise important questions regarding our understanding of PLMs using text patterns.

Numeracy enhances the Literacy of Language Models
Avijit Thawani, Jay Pujara, and Filip Ilievski 16:05–16:15

Specialized number representations in NLP have shown improvements on numerical reasoning tasks like arithmetic word problems and masked number prediction. But humans also use numeracy to make better sense of world concepts, e.g., you can seat 5 people in your 'room' but not 500. Does a better grasp of numbers improve a model's understanding of other concepts and words? This paper studies the effect of using six different number encoders on the task of masked word prediction (MWP), as a proxy for evaluating literacy. To support this investigation, we develop Wiki-Convert, a 900,000 sentence dataset annotated with numbers and units, to avoid conflating nominal and ordinal number occurrences. We find a significant improvement in MWP for sentences containing numbers, that exponent embeddings are the best number encoders, yielding over 2 points jump in prediction accuracy over a BERT baseline, and that these enhanced literacy skills also generalize to contexts without annotated numbers. We release all code at https://git.io/JuZXn.

Poster and Demo session 4

In-person Poster Session 4

Bavaro 4 *14:45–16:15*

Students Who Study Together Learn Better: On the Importance of Collective Knowledge Distillation for Domain Transfer in Fact Verification

Mitch Paul Mithun, Sandeep Suntwal, and Mihai Surdeanu

While neural networks produce state-of-the- art performance in several NLP tasks, they generally depend heavily on lexicalized information, which transfer poorly between domains. Previous works have proposed delexicalization as a form of knowledge distillation to reduce the dependency on such lexical artifacts. However, a critical unsolved issue that remains is how much delexicalization to apply: a little helps reduce overfitting, but too much discards useful information. We propose Group Learning, a knowledge and model distillation approach for fact verification in which multiple student models have access to different delexicalized views of the data, but are encouraged to learn from each other through pair-wise consistency losses. In several cross-domain experiments between the FEVER and FNC fact verification datasets, we show that our approach learns the best delexicalization strategy for the given training dataset, and outperforms state-of-the-art classifiers that rely on the original data.

MultiEURLEX - A multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer

Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos

We introduce MULTI-EURLEX, a new multilingual dataset for topic classification of legal documents. The dataset comprises 65k European Union (EU) laws, officially translated in 23 languages, annotated with multiple labels from the EUROVOC taxonomy. We highlight the effect of temporal concept drift and the importance of chronological, instead of random splits. We use the dataset as a testbed for zero-shot cross-lingual transfer, where we exploit annotated training documents in one language (source) to classify documents in another language (target). We find that fine-tuning a multilingually pretrained model (XLM-ROBERTA, MT5) in a single source language leads to catastrophic forgetting of multilingual knowledge and, consequently, poor zero-shot transfer to other languages. Adaptation strategies, namely partial fine-tuning, adapters, BITFIT, LNFIT, originally proposed to accelerate fine-tuning for new end-tasks, help retain multilingual knowledge from pretraining, substantially improving zero-shot cross-lingual transfer, but their impact also depends on the pretrained model used and the size of the label set.

Joint Passage Ranking for Diverse Multi-Answer Retrieval

Sewon Min et al.

We study multi-answer retrieval, an under-explored problem that requires retrieving passages to cover multiple distinct answers for a given question. This task requires joint modeling of retrieved passages, as models should not repeatedly retrieve passages containing the same answer at the cost of missing a different valid answer. Prior work focusing on single-answer retrieval is limited as it cannot reason about the set of passages jointly. In this paper, we introduce JPR, a joint passage retrieval model focusing on reranking. To model the joint probability of the retrieved passages, JPR makes use of an autoregressive reranker that selects a sequence of passages, equipped with novel training and decoding algorithms. Compared to prior approaches, JPR achieves significantly better answer coverage on three multi-answer datasets. When combined with downstream question answering, the improved retrieval enables larger answer generation models since they need to consider fewer passages, establishing a new state-of-the-art.

Generative Context Pair Selection for Multi-hop Question Answering

Dheeru Dua et al.

Compositional reasoning tasks such as multi-hop question answering require models to learn how to make latent decisions using only weak supervision from the final answer. Crowdsourced datasets gathered for these tasks, however, often contain only a slice of the underlying task distribution, which can induce unanticipated biases such as shallow word overlap between the question and context. Recent works have shown that discriminative training results in models that exploit these underlying biases to achieve a better held-out performance, without learning the right way to reason. We propose a generative context selection model for multi-hop QA that reasons about how the given question could have been generated given a context pair and not just independent contexts. We show that on HotpotQA, while being comparable to the state-of-the-art answering performance, our proposed generative passage selection model has a better performance (4.9% higher than baseline) on adversarial held-out set which tests robustness of model's multi-hop reasoning capabilities.

Synthetic Data Augmentation for Zero-Shot Cross-Lingual Question Answering

Arij riabi et al.

Coupled with the availability of large scale datasets, deep learning architectures have enabled rapid progress on the Question Answering task. However, most of those datasets are in English, and the performances of state-of-the-art multilingual models are significantly lower when evaluated on non-English data. Due to high data collection costs, it is not realistic to obtain annotated data for each language one desires to support. We propose a method to improve the Cross-lingual Question Answering performance without requiring additional

annotated data, leveraging Question Generation models to produce synthetic samples in a cross-lingual fashion. We show that the proposed method allows to significantly outperform the baselines trained on English data only. We report a new state-of-the-art on four datasets: MLQA, XQuAD, SQuAD-it and PIAF (fr).

Have You Seen That Number? Investigating Extrapolation in Question Answering Models
Jeonghwan Kim et al.

Numerical reasoning in machine reading comprehension (MRC) has shown drastic improvements over the past few years. While the previous models for numerical MRC are able to interpolate the learned numerical reasoning capabilities, it is not clear whether they can perform just as well on numbers unseen in the training dataset. Our work rigorously tests state-of-the-art models on DROP, a numerical MRC dataset, to see if they can handle passages that contain out-of-range numbers. One of the key findings is that the models fail to extrapolate to unseen numbers. Presenting numbers as digit-by-digit input to the model, we also propose the *E-digit* number form that alleviates the lack of extrapolation in models and reveals the need to treat numbers differently from regular words in the text. Our work provides a valuable insight into the numerical MRC models and the way to represent number forms in MRC.

Surface Form Competition: Why the Highest Probability Answer Isn't Always Right
Ari Holtzman et al.

Large language models have shown promising results in zero-shot settings. For example, they can perform multiple choice tasks simply by conditioning on a question and selecting the answer with the highest probability. However, ranking by string probability can be problematic due to surface form competition—wherein different surface forms compete for probability mass, even if they represent the same underlying concept in a given context, e.g. "computer" and "PC." Since probability mass is finite, this lowers the probability of the correct answer, due to competition from other strings that are valid answers (but not one of the multiple choice options). We introduce Domain Conditional Pointwise Mutual Information, an alternative scoring function that directly compensates for surface form competition by simply reweighing each option according to its a priori likelihood within the context of a specific task. It achieves consistent gains in zero-shot performance over both calibrated and uncalibrated scoring functions on all GPT-2 and GPT-3 models on a variety of multiple choice datasets.

Entity-Based Knowledge Conflicts in Question Answering
Shayne Longpre et al.

Knowledge-dependent tasks typically use two sources of knowledge: parametric, learned at training time, and contextual, given as a passage at inference time. To understand how models use these sources together, we formalize the problem of knowledge conflicts, where the contextual information contradicts the learned information. Analyzing the behaviour of popular models, we measure their over-reliance on memorized information (the cause of hallucinations), and uncover important factors that exacerbate this behaviour. Lastly, we propose a simple method to mitigate over-reliance on parametric knowledge, which minimizes hallucination, and improves out-of-distribution generalization by 4% - 7%. Our findings demonstrate the importance for practitioners to evaluate model tendency to hallucinate rather than read, and show that our mitigation strategy encourages generalization to evolving information (i.e. time-dependent queries). To encourage these practices, we have released our framework for generating knowledge conflicts.

Back-Training excels Self-Training at Unsupervised Domain Adaptation of Question Generation and Passage Retrieval
Devang Kulshreshtha et al.

In this work, we introduce back-training, an alternative to self-training for unsupervised domain adaptation (UDA). While self-training generates synthetic training data where natural inputs are aligned with noisy outputs, back-training results in natural outputs aligned with noisy inputs. This significantly reduces the gap between target domain and synthetic data distribution, and reduces model overfitting to source domain. We run UDA experiments on question generation and passage retrieval from the Natural Questions domain to machine learning and biomedical domains. We find that back-training vastly outperforms self-training by a mean improvement of 7.8 BLEU-4 points on generation, and 17.6% top-20 retrieval accuracy across both domains. We further propose consistency filters to remove low-quality synthetic data before training. We also release a new domain-adaptation dataset - MLQuestions containing 35K unaligned questions, 50K unaligned passages, and 3K aligned question-passage pairs.

DWUG: A large Resource of Diachronic Word Usage Graphs in Four Languages
Dominik Schlechtweg et al.

Word meaning is notoriously difficult to capture, both synchronically and diachronically. In this paper, we describe the creation of the largest resource of graded contextualized, diachronic word meaning annotation in four different languages, based on 100,000 human semantic proximity judgments. We describe in detail the multi-round incremental annotation process, the choice for a clustering algorithm to group usages into senses, and possible — diachronic and synchronic — uses for this dataset.

I Wish I Would Have Loved This One, But I Didn't – A Multilingual Dataset for Counterfactual Detection in Product Review
James O'Neill et al.

Counterfactual statements describe events that did not or cannot take place. We consider the problem of counterfactual detection (CFD) in product reviews. For this purpose, we annotate a multilingual CFD dataset from Amazon product reviews covering counterfactual statements written in English, German, and Japanese languages. The dataset is unique as it contains counterfactuals in multiple languages, covers a new application area of e-commerce reviews, and provides high quality professional annotations. We train CFD models using different text representation methods and classifiers. We find that these models are robust against the selectional biases introduced due to cue phrase-based sentence selection. Moreover, our CFD dataset is compatible with prior datasets and can be merged to learn accurate CFD models. Applying machine translation on English counterfactual examples to create multilingual data performs poorly, demonstrating the language-specificity of this problem, which has been ignored so far.

Evaluating the Morphosyntactic Well-formedness of Generated Texts

Adithya Pratapa et al.

Text generation systems are ubiquitous in natural language processing applications. However, evaluation of these systems remains a challenge, especially in multilingual settings. In this paper, we propose L'AMBRE – a metric to evaluate the morphosyntactic well-formedness of text using its dependency parse and morphosyntactic rules of the language. We present a way to automatically extract various rules governing morphosyntax directly from dependency treebanks. To tackle the noisy outputs from text generation systems, we propose a simple methodology to train robust parsers. We show the effectiveness of our metric on the task of machine translation through a diachronic study of systems translating into morphologically-rich languages.

AM2iCo: Evaluating Word Meaning in Context across Low-Resource Languages with Adversarial Examples

Qianchu Liu et al.

Capturing word meaning in context and distinguishing between correspondences and variations across languages is key to building successful multilingual and cross-lingual text representation models. However, existing multilingual evaluation datasets that evaluate lexical semantics "in-context" have various limitations. In particular, 1) their language coverage is restricted to high-resource languages and skewed in favor of only a few language families and areas, 2) a design that makes the task solvable via superficial cues, which results in artificially inflated (and sometimes super-human) performances of pretrained encoders, and 3) no support for cross-lingual evaluation. In order to address these gaps, we present AM2iCo (Adversarial and Multilingual Meaning in Context), a wide-coverage cross-lingual and multilingual evaluation set; it aims to faithfully assess the ability of state-of-the-art (SotA) representation models to understand the identity of word meaning in cross-lingual contexts for 14 language pairs. We conduct a series of experiments in a wide range of setups and demonstrate the challenging nature of AM2iCo. The results reveal that current SotA pretrained encoders substantially lag behind human performance, and the largest gaps are observed for low-resource languages and languages dissimilar to English.

Learning with Instance Bundles for Reading Comprehension

Dheeru Dua et al.

When training most modern reading comprehension models, all the questions associated with a context are treated as being independent from each other. However, closely related questions and their corresponding answers are not independent, and leveraging these relationships could provide a strong supervision signal to a model. Drawing on ideas from contrastive estimation, we introduce several new supervision losses that compare question-answer scores across multiple related instances. Specifically, we normalize these scores across various neighborhoods of closely contrasting questions and/or answers, adding a cross entropy loss term in addition to traditional maximum likelihood estimation. Our techniques require bundles of related question-answer pairs, which we either mine from within existing data or create using automated heuristics. We empirically demonstrate the effectiveness of training with instance bundles on two datasets—HotpotQA and ROPES—showing up to 9% absolute gains in accuracy.

Visually Grounded Reasoning across Languages and Cultures

Fangyu Liu et al.

The design of widespread vision-and-language datasets and pre-trained encoders directly adopts, or draws inspiration from, the concepts and images of ImageNet. While one can hardly overestimate how much this benchmark contributed to progress in computer vision, it is mostly derived from lexical databases and image queries in English, resulting in source material with a North American or Western European bias. Therefore, we devise a new protocol to construct an ImageNet-style hierarchy representative of more languages and cultures. In particular, we let the selection of both concepts and images be entirely driven by native speakers, rather than scraping them automatically. Specifically, we focus on a typologically diverse set of languages, namely, Indonesian, Mandarin Chinese, Swahili, Tamil, and Turkish. On top of the concepts and images obtained through this new protocol, we create a multilingual dataset for Multicultural Reasoning over Vision and Language (MaRVL) by eliciting statements from native speaker annotators about pairs of images. The task consists of discriminating whether each grounded statement is true or false. We establish a series of baselines using state-of-the-art models and find that their cross-lingual transfer performance lags dramatically behind supervised performance in English. These results invite us to reassess the robustness and accuracy of current state-of-the-art models beyond a narrow domain, but also open up new exciting challenges for the development of truly multilingual and multicultural systems.

CrossFit: A Few-shot Learning Challenge for Cross-task Generalization in NLP
Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren

Humans can learn a new language task efficiently with only few examples, by leveraging their knowledge obtained when learning prior tasks. In this paper, we explore whether and how such cross-task generalization ability can be acquired, and further applied to build better few-shot learners across diverse NLP tasks. We introduce CrossFit, a problem setup for studying cross-task generalization ability, which standardizes seen/unseen task partitions, data access during different learning stages, and the evaluation protocols. To instantiate different seen/unseen task partitions in CrossFit and facilitate in-depth analysis, we present the NLP Few-shot Gym, a repository of 160 diverse few-shot NLP tasks created from open-access NLP datasets and converted to a unified text-to-text format. Our analysis reveals that the few-shot learning ability on unseen tasks can be improved via an upstream learning stage using a set of seen tasks. We also observe that the selection of upstream learning tasks can significantly influence few-shot performance on unseen tasks, asking further analysis on task similarity and transferability.

On the Influence of Masking Policies in Intermediate Pre-training
Qinyuan Ye et al.

Current NLP models are predominantly trained through a two-stage "pre-train then fine-tune" pipeline. Prior work has shown that inserting an intermediate pre-training stage, using heuristic masking policies for masked language modeling (MLM), can significantly improve final performance. However, it is still unclear (1) in what cases such intermediate pre-training is helpful, (2) whether hand-crafted heuristic objectives are optimal for a given task, and (3) whether a masking policy designed for one task is generalizable beyond that task. In this paper, we perform a large-scale empirical study to investigate the effect of various masking policies in intermediate pre-training with nine selected tasks across three categories. Crucially, we introduce methods to automate the discovery of optimal masking policies via direct supervision or meta-learning. We conclude that the success of intermediate pre-training is dependent on appropriate pre-train corpus, selection of output format (i.e., masked spans or full sentence), and clear understanding of the role that MLM plays for the downstream task. In addition, we find our learned masking policies outperform the heuristic of masking named entities on TriviaQA, and policies learned from one task can positively transfer to other tasks in certain cases, inviting future research in this direction.

Back to Square One: Artifact Detection, Training and Commonsense Disentanglement in the Winograd Schema
Yanai Elazar et al.

The Winograd Schema (WS) has been proposed as a test for measuring commonsense capabilities of models. Recently, pre-trained language model-based approaches have boosted performance on some WS benchmarks but the source of improvement is still not clear. This paper suggests that the apparent progress on WS may not necessarily reflect progress in commonsense reasoning. To support this claim, we first show that the current evaluation method of WS is sub-optimal and propose a modification that uses twin sentences for evaluation. We also propose two new baselines that indicate the existence of artifacts in WS benchmarks. We then develop a method for evaluating WS-like sentences in a zero-shot setting to account for the commonsense reasoning abilities acquired during the pretraining and observe that popular language models perform randomly in this setting when using our more strict evaluation. We conclude that the observed progress is mostly due to the use of supervision in training WS models, which is not likely to successfully support all the required commonsense reasoning skills and knowledge.

ValNorm Quantifies Semantics to Reveal Consistent Valence Biases Across Languages and Over Centuries
Autumn Toney and Aylin Caliskan

Word embeddings learn implicit biases from linguistic regularities captured by word co-occurrence statistics. By extending methods that quantify human-like biases in word embeddings, we introduce ValNorm, a novel intrinsic evaluation task and method to quantify the valence dimension of affect in human-rated word sets from social psychology. We apply ValNorm on static word embeddings from seven languages (Chinese, English, German, Polish, Portuguese, Spanish, and Turkish) and from historical English text spanning 200 years. ValNorm achieves consistently high accuracy in quantifying the valence of non-discriminatory, non-social group word sets. Specifically, ValNorm achieves a Pearson correlation of r=0.88 for human judgment scores of valence for 399 words collected to establish pleasantness norms in English. In contrast, we measure gender stereotypes using the same set of word embeddings and find that social biases vary across languages. Our results indicate that valence associations of non-discriminatory, non-social group words represent widely-shared associations, in seven languages and over 200 years.

The Perils of Using Mechanical Turk to Evaluate Open-Ended Text Generation
Marzena Karpinska, Nader Akoury, and Mohit Iyyer

Recent text generation research has increasingly focused on open-ended domains such as story and poetry generation. Because models built for such tasks are difficult to evaluate automatically, most researchers in the space justify their modeling choices by collecting crowdsourced human judgments of text quality (e.g., Likert scores of coherence or grammaticality) from Amazon Mechanical Turk (AMT). In this paper, we first conduct a survey of 45 open-ended text generation papers and find that the vast majority of them fail to report crucial details about their AMT tasks, hindering reproducibility. We then run a series of story evaluation experiments with both AMT workers and English teachers and discover that even with strict qualification

filters, AMT workers (unlike teachers) fail to distinguish between model-generated text and human-generated references. We show that AMT worker judgments improve when they are shown model-generated output alongside human-generated references, which enables the workers to better calibrate their ratings. Finally, interviews with the English teachers provide deeper insights into the challenges of the evaluation process, particularly when rating model-generated text.

Perturbation CheckLists for Evaluating NLG Evaluation Metrics

Ananya B. Sai et al.

Natural Language Generation (NLG) evaluation is a multifaceted task requiring assessment of multiple desirable criteria, e.g., fluency, coherency, coverage, relevance, adequacy, overall quality, etc. Across existing datasets for 6 NLG tasks, we observe that the human evaluation scores on these multiple criteria are often not correlated. For example, there is a very low correlation between human scores on fluency and data coverage for the task of structured data to text generation. This suggests that the current recipe of proposing new automatic evaluation metrics for NLG by showing that they correlate well with scores assigned by humans for a single criteria (overall quality) alone is inadequate. Indeed, our extensive study involving 25 automatic evaluation metrics across 6 different tasks and 18 different evaluation criteria shows that there is no single metric which correlates well with human scores on all desirable criteria, for most NLG tasks. Given this situation, we propose CheckLists for better design and evaluation of automatic metrics. We design templates which target a specific criteria (e.g., coverage) and perturb the output such that the quality gets affected only along this specific criteria (e.g., the coverage drops). We show that existing evaluation metrics are not robust against even such simple perturbations and disagree with scores assigned by humans to the perturbed output. The proposed templates thus allow for a fine-grained assessment of automatic evaluation metrics exposing their limitations and will facilitate better design, analysis and evaluation of such metrics. Our templates and code are available at https://iitmnlp.github.io/EvalEval/

In-person Demo Session

Bavaro 4 *14:45–16:15*

SPRING Goes Online: End-to-End AMR Parsing and Generation

Rexhina Blloshmi et al.

In this paper we present SPRING Online Services, a Web interface and RESTful APIs for our state-of-the-art AMR parsing and generation system, SPRING (Symmetric PaRsIng aNd Generation). The Web interface has been developed to be easily used by the Natural Language Processing community, as well as by the general public. It provides, among other things, a highly interactive visualization platform and a feedback mechanism to obtain user suggestions for further improvements of the system's output. Moreover, our RESTful APIs enable easy integration of SPRING in downstream applications where AMR structures are needed. Finally, we make SPRING Online Services freely available at http://nlp.uniroma1.it/spring and, in addition, we release extra model checkpoints to be used with the original SPRING Python code.

IrEne-viz: Visualizing Energy Consumption of Transformer Models

Yash Kumar Lal et al.

IrEne is an energy prediction system that accurately predicts the interpretable inference energy consumption of a wide range of Transformer-based NLP models. We present the IrEne-viz tool, an online platform for visualizing and exploring energy consumption of various Transformer-based models easily. Additionally, we release a public API that can be used to access granular information about energy consumption of transformer models and their components. The live demo is available at http://stonybrooknlp.github.io/irene/demo/.

Session 7 Overview – Monday, November 8, 2021

	Track A	Track B	Track C	Track D	Track E	Track F
	Machine Translation and Multilinguality 3	*Question Answering 2*	*Dialogue and Interactive Systems 4*	*Resources and Evaluation 3*	*Efficient Methods for NLP 2*	*Semantics 3*
4:45	Robust Open-Vocabulary Translation from Visual Text Representations *Salesky, Etter, and Post*	How much coffee was consumed during EMNLP 2019? Fermi Problems: A New Reasoning Challenge for AI *Kalyan et al.*	ConvAbuse: Data, Analysis, and Benchmarks for Nuanced Detection in Conversational AI *Cercas Curry, Abercrombie, and Rieser*	MS^2: Multi-Document Summarization of Medical Studies *DeYoung et al.*	MATE: Multi-view Attention for Table Transformer Efficiency *Eisenschlos et al.*	Controllable Semantic Parsing via Retrieval Augmentation *Pasupat, Zhang, and Guu*
5:00	Don't Go Far Off: An Empirical Study on Neural Poetry Translation *Chakrabarty, Saakyan, and Muresan*	Will this Question be Answered? Question Filtering via Answer Model Distillation for Efficient Question Answering *Garg and Moschitti*	Conversational Multi-Hop Reasoning with Neural Commonsense Knowledge and Symbolic Logic Rules *Arabshahi et al.*	CLIPScore: A Reference-free Evaluation Metric for Image Captioning *Hessel et al.*	Learning with Different Amounts of Annotation: From Zero to Many Labels *Zhang, Gong, and Choi*	Constrained Language Models Yield Few-Shot Semantic Parsers *Shin et al.*
5:15	Improving Multilingual Translation by Representation and Gradient Regularization *Yang et al.*	Learning with Instance Bundles for Reading Comprehension *Dua et al.*	Towards Automatic Evaluation of Dialog Systems: A Model-Free Off-Policy Evaluation Approach *Jiang et al.*	On the Challenges of Evaluating Compositional Explanations in Multi-Hop Inference: Relevance, Completeness, and Expert Ratings *Jansen et al.*	When Attention Meets Fast Recurrence: Training Language Models with Reduced Compute *Lei*	ExplaGraphs: An Explanation Graph Generation Task for Structured Commonsense Reasoning *Saha et al.*
5:30	Learning Kernel-Smoothed Machine Translation with Retrieved Examples *Jiang et al.*	Explaining Answers with Entailment Trees *Dalvi et al.*	Continual Learning in Task-Oriented Dialogue Systems *Madotto et al.*	ESTER: A Machine Reading Comprehension Dataset for Reasoning about Event Semantic Relations *Han et al.*	Universal-KD: Attention-based Output-Grounded Intermediate Layer Knowledge Distillation *Wu et al.*	Connect-the-Dots: Bridging Semantics between Words and Definitions via Aligning Word Sense Inventories *Yao et al.*

Track A	**Track B**	**Track C**	**Track D**	**Track E**	**Track F**	
Machine Translation and Multilinguality 3	*Question Answering 2*	*Dialogue and Interactive Systems 4*	*Resources and Evaluation 3*	*Efficient Methods for NLP 2*	*Semantics 3*	
Uncertainty-Aware Balancing for Multilingual and Multi-Domain Neural Machine Translation Training *Wu et al.*	SituatedQA: Incorporating Extra-Linguistic Contexts into QA *Zhang and Choi*	Multilingual and Cross-Lingual Intent Detection from Spoken Data *Gerz et al.*	RICA: Evaluating Robust Inference Capabilities Based on Commonsense Axioms *Zhou et al.*	Highly Parallel Autoregressive Entity Linking with Discriminative Correction *De Cao, Aziz, and Titov*	LM-Critic: Language Models for Unsupervised Grammatical Error Correction *Yasunaga, Leskovec, and Liang*	5:45
Universal Simultaneous Machine Translation with Mixture-of-Experts Wait-k Policy *Zhang and Feng*	MKQA: A Linguistically Diverse Benchmark for Multilingual Open Domain Question Answering *Longpre, Lu, and Daiber*	Investigating Robustness of Dialog Models to Popular Figurative Language Constructs *Jhamtani et al.*	Compression, Transduction, and Creation: A Unified Framework for Evaluating Natural Language Generation *Deng et al.*	Word-Level Coreference Resolution *Dobrovolskii*	Language-agnostic Representation from Multilingual Sentence Encoders for Cross-lingual Similarity Estimation *Tiyajamorn et al.*	6:00
		Effective Sequence-to-Sequence Dialogue State Tracking *Zhao et al.*		A Secure and Efficient Federated Learning Framework for NLP *Wang et al.*		6:05

Parallel Session 7

Session 7A: Machine Translation and Multilinguality 3

Bavaro 1 *Chair: Orhan Firat*

Robust Open-Vocabulary Translation from Visual Text Representations

Elizabeth Salesky, David Etter, and Matt Post 16:45–17:00

Machine translation models have discrete vocabularies and commonly use subword segmentation techniques to achieve an 'open vocabulary.' This approach relies on consistent and correct underlying unicode sequences, and makes models susceptible to degradation from common types of noise and variation. Motivated by the robustness of human language processing, we propose the use of visual text representations, which dispense with a finite set of text embeddings in favor of continuous vocabularies created by processing visually rendered text with sliding windows. We show that models using visual text representations approach or match performance of traditional text models on small and larger datasets. More importantly, models with visual embeddings demonstrate significant robustness to varied types of noise, achieving e.g., 25.9 BLEU on a character permuted German–English task where subword models degrade to 1.9.

Don't Go Far Off: An Empirical Study on Neural Poetry Translation

Tuhin Chakrabarty, Arkadiy Saakyan, and Smaranda Muresan 17:00–17:15

Despite constant improvements in machine translation quality, automatic poetry translation remains a challenging problem due to the lack of open-sourced parallel poetic corpora, and to the intrinsic complexities involved in preserving the semantics, style and figurative nature of poetry. We present an empirical investigation for poetry translation along several dimensions: 1) size and style of training data (poetic vs. non-poetic), including a zero-shot setup; 2) bilingual vs. multilingual learning; and 3) language-family-specific models vs. mixed-language-family models. To accomplish this, we contribute a parallel dataset of poetry translations for several language pairs. Our results show that multilingual fine-tuning on poetic text significantly outperforms multilingual fine-tuning on non-poetic text that is 35X larger in size, both in terms of automatic metrics (BLEU, BERTScore, COMET) and human evaluation metrics such as faithfulness (meaning and poetic style). Moreover, multilingual fine-tuning on poetic data outperforms bilingual fine-tuning on poetic data.

Improving Multilingual Translation by Representation and Gradient Regularization

Yilin Yang et al. 17:15–17:30

Multilingual Neural Machine Translation (NMT) enables one model to serve all translation directions, including ones that are unseen during training, i.e. zero-shot translation. Despite being theoretically attractive, current models often produce low quality translations – commonly failing to even produce outputs in the right target language. In this work, we observe that off-target translation is dominant even in strong multilingual systems, trained on massive multilingual corpora. To address this issue, we propose a joint approach to regularize NMT models at both representation-level and gradient-level. At the representation level, we leverage an auxiliary target language prediction task to regularize decoder outputs to retain information about the target language. At the gradient level, we leverage a small amount of direct data (in thousands of sentence pairs) to regularize model gradients. Our results demonstrate that our approach is highly effective in both reducing off-target translation occurrences and improving zero-shot translation performance by +5.59 and +10.38 BLEU on WMT and OPUS datasets respectively. Moreover, experiments show that our method also works well when the small amount of direct data is not available.

Learning Kernel-Smoothed Machine Translation with Retrieved Examples

Qingnan Jiang et al. 17:30–17:45

How to effectively adapt neural machine translation (NMT) models according to emerging cases without retraining? Despite the great success of neural machine translation, updating the deployed models online remains a challenge. Existing non-parametric approaches that retrieve similar examples from a database to guide the translation process are promising but are prone to overfit the retrieved examples. However, non-parametric methods are prone to overfit the retrieved examples. In this work, we propose to learn Kernel-Smoothed Translation with Example Retrieval (KSTER), an effective approach to adapt neural machine translation models online. Experiments on domain adaptation and multi-domain machine translation datasets show that even without expensive retraining, KSTER is able to achieve improvement of 1.1 to 1.5 BLEU scores over the best existing online adaptation methods. The code and trained models are released at https://github.com/jiangqn/KSTER.

Uncertainty-Aware Balancing for Multilingual and Multi-Domain Neural Machine Translation Training

Minghao Wu et al. 17:45–18:00

Learning multilingual and multi-domain translation model is challenging as the heterogeneous and imbalanced data make the model converge inconsistently over different corpora in real world. One common practice is to adjust the share of each corpus in the training, so that the learning process is balanced and low-resource cases can benefit from the high resource ones. However, automatic balancing methods usually depend on the intra- and inter-dataset characteristics, which is usually agnostic or requires human priors. In this work, we propose an approach, MultiUAT, that dynamically adjusts the training data usage based on the model's uncertainty on a small set of trusted clean data for multi-corpus machine translation. We experiments with two classes

of uncertainty measures on multilingual (16 languages with 4 settings) and multi-domain settings (4 for in-domain and 2 for out-of-domain on English-German translation) and demonstrate our approach MultiUAT substantially outperforms its baselines, including both static and dynamic strategies. We analyze the cross-domain transfer and show the deficiency of static and similarity based methods.

Universal Simultaneous Machine Translation with Mixture-of-Experts Wait-k Policy
Shaolei Zhang and Yang Feng 18:00–18:15

Simultaneous machine translation (SiMT) generates translation before reading the entire source sentence and hence it has to trade off between translation quality and latency. To fulfill the requirements of different translation quality and latency in practical applications, the previous methods usually need to train multiple SiMT models for different latency levels, resulting in large computational costs. In this paper, we propose a universal SiMT model with Mixture-of-Experts Wait-k Policy to achieve the best translation quality under arbitrary latency with only one trained model. Specifically, our method employs multi-head attention to accomplish the mixture of experts where each head is treated as a wait-k expert with its own waiting words number, and given a test latency and source inputs, the weights of the experts are accordingly adjusted to produce the best translation. Experiments on three datasets show that our method outperforms all the strong baselines under different latency, including the state-of-the-art adaptive policy.

Session 7B: Question Answering 2

Bavaro 2 *Chair: Robin Jia*

How much coffee was consumed during EMNLP 2019? Fermi Problems: A New Reasoning Challenge for AI

Ashwin Kalyan et al. 16:45–17:00

Many real-world problems require the combined application of multiple reasoning abilities—employing suitable abstractions, commonsense knowledge, and creative synthesis of problem-solving strategies. To help advance AI systems towards such capabilities, we propose a new reasoning challenge, namely Fermi Problems (FPs), which are questions whose answers can only be approximately estimated because their precise computation is either impractical or impossible. For example, "How much would the sea level rise if all ice in the world melted?" FPs are commonly used in quizzes and interviews to bring out and evaluate the creative reasoning abilities of humans. To do the same for AI systems, we present two datasets: 1) A collection of 1k real-world FPs sourced from quizzes and olympiads; and 2) a bank of 10k synthetic FPs of intermediate complexity to serve as a sandbox for the harder real-world challenge. In addition to question-answer pairs, the datasets contain detailed solutions in the form of an executable program and supporting facts, helping in supervision and evaluation of intermediate steps. We demonstrate that even extensively fine-tuned large-scale language models perform poorly on these datasets, on average making estimates that are off by two orders of magnitude. Our contribution is thus the crystallization of several unsolved AI problems into a single, new challenge that we hope will spur further advances in building systems that can reason.

Will this Question be Answered? Question Filtering via Answer Model Distillation for Efficient Question Answering

Siddhant Garg and Alessandro Moschitti 17:00–17:15

In this paper we propose a novel approach towards improving the efficiency of Question Answering (QA) systems by filtering out questions that will not be answered by them. This is based on an interesting new finding: the answer confidence scores of state-of-the-art QA systems can be approximated well by models solely using the input question text. This enables preemptive filtering of questions that are not answered by the system due to their answer confidence scores being lower than the system threshold. Specifically, we learn Transformer-based question models by distilling Transformer-based answering models. Our experiments on three popular QA datasets and one industrial QA benchmark demonstrate the ability of our question models to approximate the Precision/Recall curves of the target QA system well. These question models, when used as filters, can effectively trade off lower computation cost of QA systems for lower Recall, e.g., reducing computation by ~60%, while only losing ~3-4% of Recall.

Learning with Instance Bundles for Reading Comprehension

Dheeru Dua et al. 17:15–17:30

When training most modern reading comprehension models, all the questions associated with a context are treated as being independent from each other. However, closely related questions and their corresponding answers are not independent, and leveraging these relationships could provide a strong supervision signal to a model. Drawing on ideas from contrastive estimation, we introduce several new supervision losses that compare question-answer scores across multiple related instances. Specifically, we normalize these scores across various neighborhoods of closely contrasting questions and/or answers, adding a cross entropy loss term in addition to traditional maximum likelihood estimation. Our techniques require bundles of related question-answer pairs, which we either mine from within existing data or create using automated heuristics. We empirically demonstrate the effectiveness of training with instance bundles on two datasets—HotpotQA and ROPES—showing up to 9% absolute gains in accuracy.

Explaining Answers with Entailment Trees

Bhavana Dalvi et al. 17:30–17:45

Our goal, in the context of open-domain textual question-answering (QA), is to explain answers by showing the line of reasoning from what is known to the answer, rather than simply showing a fragment of textual evidence (a "rationale"). If this could be done, new opportunities for understanding and debugging the system's reasoning become possible. Our approach is to generate explanations in the form of entailment trees, namely a tree of multipremise entailment steps from facts that are known, through intermediate conclusions, to the hypothesis of interest (namely the question + answer). To train a model with this skill, we created ENTAILMENTBANK, the first dataset to contain multistep entailment trees. Given a hypothesis (question + answer), we define three increasingly difficult explanation tasks: generate a valid entailment tree given (a) all relevant sentences (b) all relevant and some irrelevant sentences, or (c) a corpus. We show that a strong language model can partially solve these tasks, in particular when the relevant sentences are included in the input (e.g., 35% of trees for (a) are perfect), and with indications of generalization to other domains. This work is significant as it provides a new type of dataset (multistep entailments) and baselines, offering a new avenue for the community to generate richer, more systematic explanations.

SituatedQA: Incorporating Extra-Linguistic Contexts into QA

Michael Zhang and Eunsol Choi 17:45–18:00

Answers to the same question may change depending on the extra-linguistic contexts (when and where the question was asked). To study this challenge, we introduce SituatedQA, an open-retrieval QA dataset where

systems must produce the correct answer to a question given the temporal or geographical context. To construct SituatedQA, we first identify such questions in existing QA datasets. We find that a significant proportion of information seeking questions have context-dependent answers (e.g. roughly 16.5% of NQ-Open). For such context-dependent questions, we then crowdsource alternative contexts and their corresponding answers. Our study shows that existing models struggle with producing answers that are frequently updated or from uncommon locations. We further quantify how existing models, which are trained on data collected in the past, fail to generalize to answering questions asked in the present, even when provided with an updated evidence corpus (a roughly 15 point drop in accuracy). Our analysis suggests that open-retrieval QA benchmarks should incorporate extra-linguistic context to stay relevant globally and in the future. Our data, code, and datasheet are available at https://situatedqa.github.io/.

MKQA: A Linguistically Diverse Benchmark for Multilingual Open Domain Question Answering

Shayne Longpre, Yi Lu, and Joachim Daiber 18:00–18:15

Progress in cross-lingual modeling depends on challenging, realistic, and diverse evaluation sets. We introduce Multilingual Knowledge Questions and Answers (MKQA), an open-domain question answering evaluation set comprising 10k question-answer pairs aligned across 26 typologically diverse languages (260k question-answer pairs in total). The goal of this dataset is to provide a challenging benchmark for question answering quality across a wide set of languages. Answers are based on a language-independent data representation, making results comparable across languages and independent of language-specific passages. With 26 languages, this dataset supplies the widest range of languages to-date for evaluating question answering. We benchmark a wide variety of state-of-the-art methods in retrieval with generative and extractive question answering baselines, trained on Natural Questions, in zero shot and translation settings. Results indicate this dataset is challenging in all languages, and especially in low-resource languages.

Session 7C: Dialogue and Interactive Systems 4

Bavaro 3 *Chair: Emily Dinan*

ConvAbuse: Data, Analysis, and Benchmarks for Nuanced Detection in Conversational AI

Amanda Cercas Curry, Gavin Abercrombie, and Verena Rieser 16:45–17:00

We present the first English corpus study on abusive language towards three conversational AI systems gathered 'in the wild': an open-domain social bot, a rule-based chatbot, and a task-based system. To account for the complexity of the task, we take a more 'nuanced' approach where our ConvAI dataset reflects fine-grained notions of abuse, as well as views from multiple expert annotators. We find that the distribution of abuse is vastly different compared to other commonly used datasets, with more sexually tinted aggression towards the virtual persona of these systems. Finally, we report results from bench-marking existing models against this data. Unsurprisingly, we find that there is substantial room for improvement with F1 scores below 90%.

Conversational Multi-Hop Reasoning with Neural Commonsense Knowledge and Symbolic Logic Rules

Forough Arabshahi et al. 17:00–17:15

One of the challenges faced by conversational agents is their inability to identify unstated presumptions of their users' commands, a task trivial for humans due to their common sense. In this paper, we propose a zero-shot commonsense reasoning system for conversational agents in an attempt to achieve this. Our reasoner uncovers unstated presumptions from user commands satisfying a general template of if-(state), then-(action), because-(goal). Our reasoner uses a state-of-the-art transformer-based generative commonsense knowledge base (KB) as its source of background knowledge for reasoning. We propose a novel and iterative knowledge query mechanism to extract multi-hop reasoning chains from the neural KB which uses symbolic logic rules to significantly reduce the search space. Similar to any KBs gathered to date, our commonsense KB is prone to missing knowledge. Therefore, we propose to conversationally elicit the missing knowledge from human users with our novel dynamic question generation strategy, which generates and presents contextualized queries to human users. We evaluate the model with a user study with human users that achieves a 35% higher success rate compared to SOTA.

Towards Automatic Evaluation of Dialog Systems: A Model-Free Off-Policy Evaluation Approach

Haoming Jiang et al. 17:15–17:30

Reliable automatic evaluation of dialogue systems under an interactive environment has long been overdue. An ideal environment for evaluating dialog systems, also known as the Turing test, needs to involve human interaction, which is usually not affordable for large-scale experiments. Though researchers have attempted to use metrics for language generation tasks (e.g., perplexity, BLEU) or some model-based reinforcement learning methods (e.g., self-play evaluation) for automatic evaluation, these methods only show very weak correlation with the actual human evaluation in practice. To bridge such a gap, we propose a new framework named ENIGMA for estimating human evaluation scores based on recent advances of off-policy evaluation in reinforcement learning. ENIGMA only requires a handful of pre-collected experience data, and therefore does not involve human interaction with the target policy during the evaluation, making automatic evaluations feasible. More importantly, ENIGMA is model-free and agnostic to the behavior policies for collecting the experience data, which significantly alleviates the technical difficulties of modeling complex dialogue environments and human behaviors. Our experiments show that ENIGMA significantly outperforms existing methods in terms of correlation with human evaluation scores.

Continual Learning in Task-Oriented Dialogue Systems

Andrea Madotto et al. 17:30–17:45

Continual learning in task-oriented dialogue systems allows the system to add new domains and functionalities overtime after deployment, without incurring the high cost of retraining the whole system each time. In this paper, we propose a first-ever continual learning benchmark for task-oriented dialogue systems with 37 domains to be learned continuously in both modularized and end-to-end learning settings. In addition, we implement and compare multiple existing continual learning baselines, and we propose a simple yet effective architectural method based on residual adapters. We also suggest that the upper bound performance of continual learning should be equivalent to multitask learning when data from all domain is available at once. Our experiments demonstrate that the proposed architectural method and a simple replay-based strategy perform better, by a large margin, compared to other continuous learning techniques, and only slightly worse than the multitask learning upper bound while being 20X faster in learning new domains. We also report several trade-offs in terms of parameter usage, memory size and training time, which are important in the design of a task-oriented dialogue system. The proposed benchmark is released to promote more research in this direction.

Multilingual and Cross-Lingual Intent Detection from Spoken Data

Daniela Gerz et al. 17:45–17:55

We present a systematic study on multilingual and cross-lingual intent detection (ID) from spoken data. The study leverages a new resource put forth in this work, termed MInDS-14, a first training and evaluation resource for the ID task with spoken data. It covers 14 intents extracted from a commercial system in the e-banking domain, associated with spoken examples in 14 diverse language varieties. Our key results indicate that combining machine translation models with state-of-the-art multilingual sentence encoders (e.g., LaBSE) yield

strong intent detectors in the majority of target languages covered in MInDS-14, and offer comparative analyses across different axes: e.g., translation direction, impact of speech recognition, data augmentation from a related domain. We see this work as an important step towards more inclusive development and evaluation of multilingual ID from spoken data, hopefully in a much wider spectrum of languages compared to prior work.

Investigating Robustness of Dialog Models to Popular Figurative Language Constructs
Harsh Jhamtani et al. 17:55–18:05

Humans often employ figurative language use in communication, including during interactions with dialog systems. Thus, it is important for real-world dialog systems to be able to handle popular figurative language constructs like metaphor and simile. In this work, we analyze the performance of existing dialog models in situations where the input dialog context exhibits use of figurative language. We observe large gaps in handling of figurative language when evaluating the models on two open domain dialog datasets. When faced with dialog contexts consisting of figurative language, some models show very large drops in performance compared to contexts without figurative language. We encourage future research in dialog modeling to separately analyze and report results on figurative language in order to better test model capabilities relevant to real-world use. Finally, we propose lightweight solutions to help existing models become more robust to figurative language by simply using an external resource to translate figurative language to literal (non-figurative) forms while preserving the meaning to the best extent possible.

Effective Sequence-to-Sequence Dialogue State Tracking
Jeffrey Zhao et al. 18:05–18:15

Sequence-to-sequence models have been applied to a wide variety of NLP tasks, but how to properly use them for dialogue state tracking has not been systematically investigated. In this paper, we study this problem from the perspectives of pre-training objectives as well as the formats of context representations. We demonstrate that the choice of pre-training objective makes a significant difference to the state tracking quality. In particular, we find that masked span prediction is more effective than auto-regressive language modeling. We also explore using Pegasus, a span prediction-based pre-training objective for text summarization, for the state tracking model. We found that pre-training for the seemingly distant summarization task works surprisingly well for dialogue state tracking. In addition, we found that while recurrent state context representation works also reasonably well, the model may have a hard time recovering from earlier mistakes. We conducted experiments on the MultiWOZ 2.1-2.4, WOZ 2.0, and DSTC2 datasets with consistent observations.

Session 7D: Resources and Evaluation 3

Punta Cana 1,2 *Chair: Ajay Nagesh*

MS^2: Multi-Document Summarization of Medical Studies

Jay DeYoung et al. 16:45–17:00

To assess the effectiveness of any medical intervention, researchers must conduct a time-intensive and manual literature review. NLP systems can help to automate or assist in parts of this expensive process. In support of this goal, we release MS^2 (Multi-Document Summarization of Medical Studies), a dataset of over 470k documents and 20K summaries derived from the scientific literature. This dataset facilitates the development of systems that can assess and aggregate contradictory evidence across multiple studies, and is the first large-scale, publicly available multi-document summarization dataset in the biomedical domain. We experiment with a summarization system based on BART, with promising early results, though significant work remains to achieve higher summarization quality. We formulate our summarization inputs and targets in both free text and structured forms and modify a recently proposed metric to assess the quality of our system's generated summaries. Data and models are available at https://github.com/allenai/ms2.

CLIPScore: A Reference-free Evaluation Metric for Image Captioning

Jack Hessel et al. 17:00–17:15

Image captioning has conventionally relied on reference-based automatic evaluations, where machine captions are compared against captions written by humans. This is in contrast to the reference-free manner in which humans assess caption quality. In this paper, we report the surprising empirical finding that CLIP (Radford et al., 2021), a cross-modal model pretrained on 400M image+caption pairs from the web, can be used for robust automatic evaluation of image captioning without the need for references. Experiments spanning several corpora demonstrate that our new reference-free metric, CLIPScore, achieves the highest correlation with human judgements, outperforming existing reference-based metrics like CIDEr and SPICE. Information gain experiments demonstrate that CLIPScore, with its tight focus on image-text compatibility, is complementary to existing reference-based metrics that emphasize text-text similarities. Thus, we also present a reference-augmented version, RefCLIPScore, which achieves even higher correlation. Beyond literal description tasks, several case studies reveal domains where CLIPScore performs well (clip-art images, alt-text rating), but also where it is relatively weaker in comparison to reference-based metrics, e.g., news captions that require richer contextual knowledge.

On the Challenges of Evaluating Compositional Explanations in Multi-Hop Inference: Relevance, Completeness, and Expert Ratings

Peter Jansen et al. 17:15–17:30

Building compositional explanations requires models to combine two or more facts that, together, describe why the answer to a question is correct. Typically, these "multi-hop" explanations are evaluated relative to one (or a small number of) gold explanations. In this work, we show these evaluations substantially underestimate model performance, both in terms of the relevance of included facts, as well as the completeness of model-generated explanations, because models regularly discover and produce valid explanations that are different than gold explanations. To address this, we construct a large corpus of 126k domain-expert (science teacher) relevance ratings that augment a corpus of explanations to standardized science exam questions, discovering 80k additional relevant facts not rated as gold. We build three strong models based on different methodologies (generation, ranking, and schemas), and empirically show that while expert-augmented ratings provide better estimates of explanation quality, both original (gold) and expert-augmented automatic evaluations still substantially underestimate performance by up to 36% when compared with full manual expert judgements, with different models being disproportionately affected. This poses a significant methodological challenge to accurately evaluating explanations produced by compositional reasoning models.

ESTER: A Machine Reading Comprehension Dataset for Reasoning about Event Semantic Relations

Rujun Han et al. 17:30–17:45

Understanding how events are semantically related to each other is the essence of reading comprehension. Recent event-centric reading comprehension datasets focus mostly on event arguments or temporal relations. While these tasks partially evaluate machines' ability of narrative understanding, human-like reading comprehension requires the capability to process event-based information beyond arguments and temporal reasoning. For example, to understand causality between events, we need to infer motivation or purpose; to establish event hierarchy, we need to understand the composition of events. To facilitate these tasks, we introduce **ESTER**, a comprehensive machine reading comprehension (MRC) dataset for Event Semantic Relation Reasoning. The dataset leverages natural language queries to reason about the five most common event semantic relations, provides more than 6K questions, and captures 10.1K event relation pairs. Experimental results show that the current SOTA systems achieve 22.1%, 63.3% and 83.5% for token-based exact-match (**EM**), **F1** and event-based **HIT1** scores, which are all significantly below human performances (36.0%, 79.6%, 100% respectively), highlighting our dataset as a challenging benchmark.

RICA: Evaluating Robust Inference Capabilities Based on Commonsense Axioms

Pei Zhou et al. 17:45–18:00

Pre-trained language models (PTLMs) have achieved impressive performance on commonsense inference benchmarks, but their ability to employ commonsense to make robust inferences, which is crucial for effective communications with humans, is debated. In the pursuit of advancing fluid human-AI communication, we propose a new challenge, RICA: Robust Inference using Commonsense Axioms, that evaluates robust commonsense inference despite textual perturbations. To generate data for this challenge, we develop a systematic and scalable procedure using commonsense knowledge bases and probe PTLMs across two different evaluation settings. Extensive experiments on our generated probe sets with more than 10k statements show that PTLMs perform no better than random guessing on the zero-shot setting, are heavily impacted by statistical biases, and are not robust to perturbation attacks. We also find that fine-tuning on similar statements offer limited gains, as PTLMs still fail to generalize to unseen inferences. Our new large-scale benchmark exposes a significant gap between PTLMs and human-level language understanding and offers a new challenge for PTLMs to demonstrate commonsense.

Compression, Transduction, and Creation: A Unified Framework for Evaluating Natural Language Generation

Mingkai Deng et al. 18:00–18:15

Natural language generation (NLG) spans a broad range of tasks, each of which serves for specific objectives and desires different properties of generated text. The complexity makes automatic evaluation of NLG particularly challenging. Previous work has typically focused on a single task and developed individual evaluation metrics based on specific intuitions. In this paper, we propose a unifying perspective based on the nature of information change in NLG tasks, including compression (e.g., summarization), transduction (e.g., text rewriting), and creation (e.g., dialog). _Information alignment_ between input, context, and output text plays a common central role in characterizing the generation. With automatic alignment prediction models, we develop a family of interpretable metrics that are suitable for evaluating key aspects of different NLG tasks, often without need of gold reference data. Experiments show the uniformly designed metrics achieve stronger or comparable correlations with human judgement compared to state-of-the-art metrics in each of diverse tasks, including text summarization, style transfer, and knowledge-grounded dialog.

Session 7E: Efficient Methods for NLP 2

Punta Cana 3,4 *Chair: Jesse Dodge*

MATE: Multi-view Attention for Table Transformer Efficiency

Julian Eisenschlos et al. 16:45–17:00

This work presents a sparse-attention Transformer architecture for modeling documents that contain large tables. Tables are ubiquitous on the web, and are rich in information. However, more than 20% of relational tables on the web have 20 or more rows (Cafarella et al., 2008), and these large tables present a challenge for current Transformer models, which are typically limited to 512 tokens. Here we propose MATE, a novel Transformer architecture designed to model the structure of web tables. MATE uses sparse attention in a way that allows heads to efficiently attend to either rows or columns in a table. This architecture scales linearly with respect to speed and memory, and can handle documents containing more than 8000 tokens with current accelerators. MATE also has a more appropriate inductive bias for tabular data, and sets a new state-of-the-art for three table reasoning datasets. For HybridQA (Chen et al., 2020), a dataset that involves large documents containing tables, we improve the best prior result by 19 points.

Learning with Different Amounts of Annotation: From Zero to Many Labels

Shujian Zhang, Chengyue Gong, and Eunsol Choi 17:00–17:15

Training NLP systems typically assumes access to annotated data that has a single human label per example. Given imperfect labeling from annotators and inherent ambiguity of language, we hypothesize that single label is not sufficient to learn the spectrum of language interpretation. We explore new annotation distribution schemes, assigning multiple labels per example for a small subset of training examples. Introducing such multi label examples at the cost of annotating fewer examples brings clear gains on natural language inference task and entity typing task, even when we simply first train with a single label data and then fine tune with multi label examples. Extending a MixUp data augmentation framework, we propose a learning algorithm that can learn from training examples with different amount of annotation (with zero, one, or multiple labels). This algorithm efficiently combines signals from uneven training data and brings additional gains in low annotation budget and cross domain settings. Together, our method achieves consistent gains in two tasks, suggesting distributing labels unevenly among training examples can be beneficial for many NLP tasks.

When Attention Meets Fast Recurrence: Training Language Models with Reduced Compute

Tao Lei 17:15–17:30

Large language models have become increasingly difficult to train because of the growing computation time and cost. In this work, we present SRU++, a highly-efficient architecture that combines fast recurrence and attention for sequence modeling. SRU++ exhibits strong modeling capacity and training efficiency. On standard language modeling tasks such as Enwik8, Wiki-103 and Billion Word datasets, our model obtains better bits-per-character and perplexity while using 3x-10x less training cost compared to top-performing Transformer models. For instance, our model achieves a state-of-the-art result on the Enwik8 dataset using 1.6 days of training on an 8-GPU machine. We further demonstrate that SRU++ requires minimal attention for near state-of-the-art performance. Our results suggest jointly leveraging fast recurrence with little attention as a promising direction for accelerating model training and inference.

Universal-KD: Attention-based Output-Grounded Intermediate Layer Knowledge Distillation

Yimeng Wu et al. 17:30–17:45

Intermediate layer matching is shown as an effective approach for improving knowledge distillation (KD). However, this technique applies matching in the hidden spaces of two different networks (i.e. student and teacher), which lacks clear interpretability. Moreover, intermediate layer KD cannot easily deal with other problems such as layer mapping search and architecture mismatch (i.e. it requires the teacher and student to be of the same model type). To tackle the aforementioned problems all together, we propose Universal-KD to match intermediate layers of the teacher and the student in the output space (by adding pseudo classifiers on intermediate layers) via the attention-based layer projection. By doing this, our unified approach has three merits: (i) it can be flexibly combined with current intermediate layer distillation techniques to improve their results (ii) the pseudo classifiers of the teacher can be deployed instead of extra expensive teacher assistant networks to address the capacity gap problem in KD which is a common issue when the gap between the size of the teacher and student networks becomes too large; (iii) it can be used in cross-architecture intermediate layer KD. We did comprehensive experiments in distilling BERT-base into BERT-4, RoBERTa-large into DistilRoBERTa and BERT-base into CNN and LSTM-based models. Results on the GLUE tasks show that our approach is able to outperform other KD techniques.

Highly Parallel Autoregressive Entity Linking with Discriminative Correction

Nicola De Cao, Wilker Aziz, and Ivan Titov 17:45–17:55

Generative approaches have been recently shown to be effective for both Entity Disambiguation and Entity Linking (i.e., joint mention detection and disambiguation). However, the previously proposed autoregressive formulation for EL suffers from i) high computational cost due to a complex (deep) decoder, ii) non-parallelizable decoding that scales with the source sequence length, and iii) the need for training on a large amount of data. In this work, we propose a very efficient approach that parallelizes autoregressive linking

across all potential mentions and relies on a shallow and efficient decoder. Moreover, we augment the generative objective with an extra discriminative component, i.e., a correction term which lets us directly optimize the generator's ranking. When taken together, these techniques tackle all the above issues: our model is >70 times faster and more accurate than the previous generative method, outperforming state-of-the-art approaches on the standard English dataset AIDA-CoNLL. Source code available at https://github.com/nicola-decao/efficient-autoregressive-EL

Word-Level Coreference Resolution
Vladimir Dobrovolskii 17:55–18:05

Recent coreference resolution models rely heavily on span representations to find coreference links between word spans. As the number of spans is $O(n^2)$ in the length of text and the number of potential links is $O(n^4)$, various pruning techniques are necessary to make this approach computationally feasible. We propose instead to consider coreference links between individual words rather than word spans and then reconstruct the word spans. This reduces the complexity of the coreference model to $O(n^2)$ and allows it to consider all potential mentions without pruning any of them out. We also demonstrate that, with these changes, SpanBERT for coreference resolution will be significantly outperformed by RoBERTa. While being highly efficient, our model performs competitively with recent coreference resolution systems on the OntoNotes benchmark.

A Secure and Efficient Federated Learning Framework for NLP
Chenghong Wang et al. 18:05–18:15

In this work, we consider the problem of designing secure and efficient federated learning (FL) frameworks for NLP. Existing solutions under this literature either consider a trusted aggregator or require heavy-weight cryptographic primitives, which makes the performance significantly degraded. Moreover, many existing secure FL designs work only under the restrictive assumption that none of the clients can be dropped out from the training protocol. To tackle these problems, we propose SEFL, a secure and efficient federated learning framework that (1)~eliminates the need for the trusted entities; (2)~achieves similar and even better model accuracy compared with existing FL designs; (3)~is resilient to client dropouts.

Session 7F: Semantics 3

Santo Domingo *Chair: Swabha Swayamdipta*

Controllable Semantic Parsing via Retrieval Augmentation

Panupong Pasupat, Yuan Zhang, and Kelvin Guu 16:45–17:00

In practical applications of semantic parsing, we often want to rapidly change the behavior of the parser, such as enabling it to handle queries in a new domain, or changing its predictions on certain targeted queries. While we can introduce new training examples exhibiting the target behavior, a mechanism for enacting such behavior changes without expensive model re-training would be preferable. To this end, we propose ControllAble Semantic Parser via Exemplar Retrieval (CASPER). Given an input query, the parser retrieves related exemplars from a retrieval index, augments them to the query, and then applies a generative seq2seq model to produce an output parse. The exemplars act as a control mechanism over the generic generative model: by manipulating the retrieval index or how the augmented query is constructed, we can manipulate the behavior of the parser. On the MTOP dataset, in addition to achieving state-of-the-art on the standard setup, we show that CASPER can parse queries in a new domain, adapt the prediction toward the specified patterns, or adapt to new semantic schemas without having to further re-train the model.

Constrained Language Models Yield Few-Shot Semantic Parsers

Richard Shin et al. 17:00–17:15

We explore the use of large pretrained language models as few-shot semantic parsers. The goal in semantic parsing is to generate a structured meaning representation given a natural language input. However, language models are trained to generate natural language. To bridge the gap, we use language models to paraphrase inputs into a controlled sublanguage resembling English that can be automatically mapped to a target meaning representation. Our results demonstrate that with only a small amount of data and very little code to convert into English-like representations, our blueprint for rapidly bootstrapping semantic parsers leads to surprisingly effective performance on multiple community tasks, greatly exceeding baseline methods also trained on the same limited data.

ExplaGraphs: An Explanation Graph Generation Task for Structured Commonsense Reasoning

Swarnadeep Saha et al. 17:15–17:30

Recent commonsense-reasoning tasks are typically discriminative in nature, where a model answers a multiple-choice question for a certain context. Discriminative tasks are limiting because they fail to adequately evaluate the model's ability to reason and explain predictions with underlying commonsense knowledge. They also allow such models to use reasoning shortcuts and not be "right for the right reasons". In this work, we present ExplaGraphs, a new generative and structured commonsense-reasoning task (and an associated dataset) of explanation graph generation for stance prediction. Specifically, given a belief and an argument, a model has to predict if the argument supports or counters the belief and also generate a commonsense-augmented graph that serves as non-trivial, complete, and unambiguous explanation for the predicted stance. We collect explanation graphs through a novel Create-Verify-And-Refine graph collection framework that improves the graph quality (up to 90%) via multiple rounds of verification and refinement. A significant 79% of our graphs contain external commonsense nodes with diverse structures and reasoning depths. Next, we propose a multi-level evaluation framework, consisting of automatic metrics and human evaluation, that check for the structural and semantic correctness of the generated graphs and their degree of match with ground-truth graphs. Finally, we present several structured, commonsense-augmented, and text generation models as strong starting points for this explanation graph generation task, and observe that there is a large gap with human performance, thereby encouraging future work for this new challenging task.

Connect-the-Dots: Bridging Semantics between Words and Definitions via Aligning Word Sense Inventories

Wenlin Yao et al. 17:30–17:45

Word Sense Disambiguation (WSD) aims to automatically identify the exact meaning of one word according to its context. Existing supervised models struggle to make correct predictions on rare word senses due to limited training data and can only select the best definition sentence from one predefined word sense inventory (e.g., WordNet). To address the data sparsity problem and generalize the model to be independent of one predefined inventory, we propose a gloss alignment algorithm that can align definition sentences (glosses) with the same meaning from different sense inventories to collect rich lexical knowledge. We then train a model to identify semantic equivalence between a target word in context and one of its glosses using these aligned inventories, which exhibits strong transfer capability to many WSD tasks. Experiments on benchmark datasets show that the proposed method improves predictions on both frequent and rare word senses, outperforming prior work by 1.2% on the All-Words WSD Task and 4.3% on the Low-Shot WSD Task. Evaluation on WiC Task also indicates that our method can better capture word meanings in context.

LM-Critic: Language Models for Unsupervised Grammatical Error Correction

Michihiro Yasunaga, Jure Leskovec, and Percy Liang 17:45–18:00

Grammatical error correction (GEC) requires a set of labeled ungrammatical / grammatical sentence pairs for training, but obtaining such annotation can be prohibitively expensive. Recently, the Break-It-Fix-It (BIFI) framework has demonstrated strong results on learning to repair a broken program without any labeled exam-

ples, but this relies on a perfect critic (e.g., a compiler) that returns whether an example is valid or not, which does not exist for the GEC task. In this work, we show how to leverage a pretrained language model (LM) in defining an LM-Critic, which judges a sentence to be grammatical if the LM assigns it a higher probability than its local perturbations. We apply this LM-Critic and BIFI along with a large set of unlabeled sentences to bootstrap realistic ungrammatical / grammatical pairs for training a corrector. We evaluate our approach on GEC datasets on multiple domains (CoNLL-2014, BEA-2019, GMEG-wiki and GMEG-yahoo) and show that it outperforms existing methods in both the unsupervised setting (+7.7 F0.5) and the supervised setting (+0.5 F0.5).

Language-agnostic Representation from Multilingual Sentence Encoders for Cross-lingual Similarity Estimation
Nattapong Tiyajamorn et al. 18:00–18:15

We propose a method to distill a language-agnostic meaning embedding from a multilingual sentence encoder. By removing language-specific information from the original embedding, we retrieve an embedding that fully represents the sentence's meaning. The proposed method relies only on parallel corpora without any human annotations. Our meaning embedding allows efficient cross-lingual sentence similarity estimation by simple cosine similarity calculation. Experimental results on both quality estimation of machine translation and cross-lingual semantic textual similarity tasks reveal that our method consistently outperforms the strong baselines using the original multilingual embedding. Our method consistently improves the performance of any pre-trained multilingual sentence encoder, even in low-resource language pairs where only tens of thousands of parallel sentence pairs are available.

Virtual Poster and Demo Session 2

Time: *12:30–14:30*

Aligning Multidimensional Worldviews and Discovering Ideological Differences
Jeremiah Milbauer, Adarsh Mathew, and James Evans

The Internet is home to thousands of communities, each with their own unique worldview and associated ideological differences. With new communities constantly emerging and serving as ideological birthplaces, battlegrounds, and bunkers, it is critical to develop a framework for understanding worldviews and ideological distinction. Most existing work, however, takes a predetermined view based on political polarization: the "right vs. left" dichotomy of U.S. politics. In reality, both political polarization – and worldviews more broadly – transcend one-dimensional difference, and deserve a more complete analysis. Extending the ability of word embedding models to capture the semantic and cultural characteristics of their training corpora, we propose a novel method for discovering the multifaceted ideological and worldview characteristics of communities. Using over 1B comments collected from the largest communities on Reddit.com representing ~40% of Reddit activity, we demonstrate the efficacy of this approach to uncover complex ideological differences across multiple axes of polarization.

Just Say No: Analyzing the Stance of Neural Dialogue Generation in Offensive Contexts
Ashutosh Baheti et al.

Dialogue models trained on human conversations inadvertently learn to generate toxic responses. In addition to producing explicitly offensive utterances, these models can also implicitly insult a group or individual by aligning themselves with an offensive statement. To better understand the dynamics of contextually offensive language, we investigate the stance of dialogue model responses in offensive Reddit conversations. Specifically, we create ToxiChat, a crowd-annotated dataset of 2,000 Reddit threads and model responses labeled with offensive language and stance. Our analysis reveals that 42% of human responses agree with toxic comments, whereas only 13% agree with safe comments. This undesirable behavior is learned by neural dialogue models, such as DialoGPT, which we show are two times more likely to agree with offensive comments. To enable automatic detection of offensive language, we fine-tuned transformer-based classifiers on ToxiChat that achieve 0.71 F1 for offensive labels and 0.53 Macro-F1 for stance labels. Finally, we quantify the effectiveness of controllable text generation (CTG) methods to mitigate the tendency of neural dialogue models to agree with offensive comments. Compared to the baseline, our best CTG model achieves a 19% reduction in agreement with offensive comments and produces 29% fewer offensive replies. Our work highlights the need for further efforts to characterize and analyze inappropriate behavior in dialogue models, in order to help make them safer.

Multi-Modal Open-Domain Dialogue
Kurt Shuster et al.

Recent work in open-domain conversational agents has demonstrated that significant improvements in humanness and user preference can be achieved via massive scaling in both pre-training data and model size (Adiwardana et al., 2020; Roller et al., 2020). However, if we want to build agents with human-like abilities, we must expand beyond handling just text. A particularly important topic is the ability to see images and communicate about what is perceived. With the goal of getting humans to engage in multi-modal dialogue, we investigate combining components from state-of-the-art open-domain dialogue agents with those from state-of-the-art vision models. We study incorporating different image fusion schemes and domain-adaptive pre-training and fine-tuning strategies, and show that our best resulting model outperforms strong existing models in multi-modal dialogue while simultaneously performing as well as its predecessor (text-only) BlenderBot (Roller et al., 2020) in text-based conversation. We additionally investigate and incorporate safety components in our final model, and show that such efforts do not diminish model performance with respect to human preference.

A Label-Aware BERT Attention Network for Zero-Shot Multi-Intent Detection in Spoken Language Understanding
Ting-Wei Wu, Ruolin Su, and Biing Juang

With the early success of query-answer assistants such as Alexa and Siri, research attempts to expand system capabilities of handling service automation are now abundant. However, preliminary systems have quickly found the inadequacy in relying on simple classification techniques to effectively accomplish the automation task. The main challenge is that the dialogue often involves complexity in user's intents (or purposes) which are multiproned, subject to spontaneous change, and difficult to track. Furthermore, public datasets have not considered these complications and the general semantic annotations are lacking which may result in zero-shot problem. Motivated by the above, we propose a Label-Aware BERT Attention Network (LABAN) for zero-shot multi-intent detection. We first encode input utterances with BERT and construct a label embedded space by considering embedded semantics in intent labels. An input utterance is then classified based on its projection weights on each intent embedding in this embedded space. We show that it successfully extends to few/zero-shot setting where part of intent labels are unseen in training data, by also taking account of semantics in these unseen intent labels. Experimental results show that our approach is capable of detecting many unseen intent labels correctly. It also achieves the state-of-the-art performance on five multi-intent datasets in normal cases.

Zero-Shot Dialogue Disentanglement by Self-Supervised Entangled Response Selection
Ta-Chung Chi

Dialogue disentanglement aims to group utterances in a long and multi-participant dialogue into threads. This is useful for discourse analysis and downstream applications such as dialogue response selection, where it can be the first step to construct a clean context/response set. Unfortunately, labeling all~*reply-to* links takes quadratic effort w.r.t the number of utterances: an annotator must check all preceding utterances to identify the one to which the current utterance is a reply. In this paper, we are the first to propose a~**zero-shot** dialogue disentanglement solution. Firstly, we train a model on a multi-participant response selection dataset harvested from the web which is not annotated; we then apply the trained model to perform zero-shot dialogue disentanglement. Without any labeled data, our model can achieve a cluster F1 score of 25. We also fine-tune the model using various amounts of labeled data. Experiments show that with only 10% of the data, we achieve nearly the same performance of using the full dataset. Code is released at `https://github.com/chijames`.

SIMMC 2.0: A Task-oriented Dialog Dataset for Immersive Multimodal Conversations
Satwik Kottur et al.

Next generation task-oriented dialog systems need to understand conversational contexts with their perceived surroundings, to effectively help users in the real-world multimodal environment. Existing task-oriented dialog datasets aimed towards virtual assistance fall short and do not situate the dialog in the user's multimodal context. To overcome, we present a new dataset for Situated and Interactive Multimodal Conversations, SIMMC 2.0, which includes 11K task-oriented user<->assistant dialogs (117K utterances) in the shopping domain, grounded in immersive and photo-realistic scenes. The dialogs are collection using a two-phase pipeline: (1) A novel multimodal dialog simulator generates simulated dialog flows, with an emphasis on diversity and richness of interactions, (2) Manual paraphrasing of generating utterances to draw from natural language distribution. We provide an in-depth analysis of the collected dataset, and describe in detail the four main benchmark tasks we propose for SIMMC 2.0. Our baseline model, powered by the state-of-the-art language model, shows promising results, and highlights new challenges and directions for the community to study.

RAST: Domain-Robust Dialogue Rewriting as Sequence Tagging
Jie Hao et al.

The task of dialogue rewriting aims to reconstruct the latest dialogue utterance by copying the missing content from the dialogue context. Until now, the existing models for this task suffer from the robustness issue, i.e., performances drop dramatically when testing on a different dataset. We address this robustness issue by proposing a novel sequence-tagging-based model so that the search space is significantly reduced, yet the core of this task is still well covered. As a common issue of most tagging models for text generation, the model's outputs may lack fluency. To alleviate this issue, we inject the loss signal from BLEU or GPT-2 under a REINFORCE framework. Experiments show huge improvements of our model over the current state-of-the-art systems when transferring to another dataset.

MRF-Chat: Improving Dialogue with Markov Random Fields
Ishaan Grover et al.

Recent state-of-the-art approaches in open-domain dialogue include training end-to-end deep-learning models to learn various conversational features like emotional content of response, symbolic transitions of dialogue contexts in a knowledge graph and persona of the agent and the user, among others. While neural models have shown reasonable results, modelling the cognitive processes that humans use when conversing with each other may improve the agent's quality of responses. A key element of natural conversation is to tailor one's response such that it accounts for concepts that the speaker and listener may or may not know and the contextual relevance of all prior concepts used in conversation. We show that a rich representation and explicit modeling of these psychological processes can improve predictions made by existing neural network models. In this work, we propose a novel probabilistic approach using Markov Random Fields (MRF) to augment existing deep-learning methods for improved next utterance prediction. Using human and automatic evaluations, we show that our augmentation approach significantly improves the performance of existing state-of-the-art retrieval models for open-domain conversational agents.

Dialogue State Tracking with a Language Model using Schema-Driven Prompting
Chia-Hsuan Lee, Hao Cheng, and Mari Ostendorf

Task-oriented conversational systems often use dialogue state tracking to represent the user's intentions, which involves filling in values of pre-defined slots. Many approaches have been proposed, often using task-specific architectures with special-purpose classifiers. Recently, good results have been obtained using more general architectures based on pretrained language models. Here, we introduce a new variation of the language modeling approach that uses schema-driven prompting to provide task-aware history encoding that is used for both categorical and non-categorical slots. We further improve performance by augmenting the prompting with schema descriptions, a naturally occurring source of in-domain knowledge. Our purely generative system achieves state-of-the-art performance on MultiWOZ 2.2 and achieves competitive performance on two other benchmarks: MultiWOZ 2.1 and M2M. The data and code will be available at https://github.com/chiahsuan156/DST-as-Prompting.

Signed Coreference Resolution
Kayo Yin, Kenneth DeHaan, and Malihe Alikhani

Coreference resolution is key to many natural language processing tasks and yet has been relatively unexplored in Sign Language Processing. In signed languages, space is primarily used to establish reference. Solving coreference resolution for signed languages would not only enable higher-level Sign Language Processing systems, but also enhance our understanding of language in different modalities and of situated references, which are key problems in studying grounded language. In this paper, we: (1) introduce Signed Coreference Resolution (SCR), a new challenge for coreference modeling and Sign Language Processing; (2) collect an annotated corpus of German Sign Language with gold labels for coreference together with an annotation software for the task; (3) explore features of hand gesture, iconicity, and spatial situated properties and move forward to propose a set of linguistically informed heuristics and unsupervised models for the task; (4) put forward several proposals about ways to address the complexities of this challenge effectively.

Consistent Accelerated Inference via Confident Adaptive Transformers
Tal Schuster et al.

We develop a novel approach for confidently accelerating inference in the large and expensive multilayer Transformers that are now ubiquitous in natural language processing (NLP). Amortized or approximate computational methods increase efficiency, but can come with unpredictable performance costs. In this work, we present CATs – Confident Adaptive Transformers – in which we simultaneously increase computational efficiency, while guaranteeing a specifiable degree of consistency with the original model with high confidence. Our method trains additional prediction heads on top of intermediate layers, and dynamically decides when to stop allocating computational effort to each input using a meta consistency classifier. To calibrate our early prediction stopping rule, we formulate a unique extension of conformal prediction. We demonstrate the effectiveness of this approach on four classification and regression tasks.

Improving and Simplifying Pattern Exploiting Training
Derek Tam et al.

Recently, pre-trained language models (LMs) have achieved strong performance when fine-tuned on difficult benchmarks like SuperGLUE. However, performance can suffer when there are very few labeled examples available for fine-tuning. Pattern Exploiting Training (PET) is a recent approach that leverages patterns for few-shot learning. However, PET uses task-specific unlabeled data. In this paper, we focus on few-shot learning without any unlabeled data and introduce ADAPET, which modifies PET's objective to provide denser supervision during fine-tuning. As a result, ADAPET outperforms PET on SuperGLUE without any task-specific unlabeled data.

Unsupervised Data Augmentation with Naive Augmentation and without Unlabeled Data
David Lowell et al.

Unsupervised Data Augmentation (UDA) is a semisupervised technique that applies a consistency loss to penalize differences between a model's predictions on (a) observed (unlabeled) examples; and (b) corresponding 'noised' examples produced via data augmentation. While UDA has gained popularity for text classification, open questions linger over which design decisions are necessary and how to extend the method to sequence labeling tasks. In this paper, we re-examine UDA and demonstrate its efficacy on several sequential tasks. Our main contribution is an empirical study of UDA to establish which components of the algorithm confer benefits in NLP. Notably, although prior work has emphasized the use of clever augmentation techniques including back-translation, we find that enforcing consistency between predictions assigned to observed and randomly substituted words often yields comparable (or greater) benefits compared to these more complex perturbation models. Furthermore, we find that applying UDA's consistency loss affords meaningful gains without any unlabeled data at all, i.e., in a standard supervised setting. In short, UDA need not be unsupervised to realize much of its noted benefits, and does not require complex data augmentation to be effective.

Pre-train or Annotate? Domain Adaptation with a Constrained Budget
Fan Bai, Alan Ritter, and Wei Xu

Recent work has demonstrated that pre-training in-domain language models can boost performance when adapting to a new domain. However, the costs associated with pre-training raise an important question: given a fixed budget, what steps should an NLP practitioner take to maximize performance? In this paper, we study domain adaptation under budget constraints, and approach it as a customer choice problem between data annotation and pre-training. Specifically, we measure the annotation cost of three procedural text datasets and the pre-training cost of three in-domain language models. Then we evaluate the utility of different combinations of pre-training and data annotation under varying budget constraints to assess which combination strategy works best. We find that, for small budgets, spending all funds on annotation leads to the best performance; once the budget becomes large enough, a combination of data annotation and in-domain pre-training works more optimally. We therefore suggest that task-specific data annotation should be part of an economical strategy when adapting an NLP model to a new domain.

Lawyers are Dishonest? Quantifying Representational Harms in Commonsense Knowledge Resources
Ninareh Mehrabi et al.

Warning: this paper contains content that may be offensive or upsetting. Commonsense knowledge bases (CSKB) are increasingly used for various natural language processing tasks. Since CSKBs are mostly human-generated and may reflect societal biases, it is important to ensure that such biases are not conflated with the notion of commonsense. Here we focus on two widely used CSKBs, ConceptNet and GenericsKB, and estab-

lish the presence of bias in the form of two types of representational harms, overgeneralization of polarized perceptions and representation disparity across different demographic groups in both CSKBs. Next, we find similar representational harms for downstream models that use ConceptNet. Finally, we propose a filtering-based approach for mitigating such harms, and observe that our filtered-based approach can reduce the issues in both resources and models but leads to a performance drop, leaving room for future work to build fairer and stronger commonsense models.

Sentence-Permuted Paragraph Generation
Wenhao Yu et al.

Generating paragraphs of diverse contents is important in many applications. Existing generation models produce similar contents from homogenized contexts due to the fixed left-to-right sentence order. Our idea is permuting the sentence orders to improve the content diversity of multi-sentence paragraph. We propose a novel framework PermGen whose objective is to maximize the expected log-likelihood of output paragraph distributions with respect to all possible sentence orders. PermGen uses hierarchical positional embedding and designs new procedures for training, and decoding in the sentence-permuted generation. Experiments on three paragraph generation benchmarks demonstrate PermGen generates more diverse outputs with a higher quality than existing models.

Extract, Denoise and Enforce: Evaluating and Improving Concept Preservation for Text-to-Text Generation
Yuning Mao et al.

Prior studies on text-to-text generation typically assume that the model could figure out what to attend to in the input and what to include in the output via seq2seq learning, with only the parallel training data and no additional guidance. However, it remains unclear whether current models can preserve important concepts in the source input, as seq2seq learning does not have explicit focus on the concepts and commonly used evaluation metrics also treat them equally important as other tokens. In this paper, we present a systematic analysis that studies whether current seq2seq models, especially pre-trained language models, are good enough for preserving important input concepts and to what extent explicitly guiding generation with the concepts as lexical constraints is beneficial. We answer the above questions by conducting extensive analytical experiments on four representative text-to-text generation tasks. Based on the observations, we then propose a simple yet effective framework to automatically extract, denoise, and enforce important input concepts as lexical constraints. This new method performs comparably or better than its unconstrained counterpart on automatic metrics, demonstrates higher coverage for concept preservation, and receives better ratings in the human evaluation. Our code is available at https://github.com/morningmoni/EDE.

Paraphrase Generation: A Survey of the State of the Art
Jianing Zhou and Suma Bhat

This paper focuses on paraphrase generation,which is a widely studied natural language generation task in NLP. With the development of neural models, paraphrase generation research has exhibited a gradual shift to neural methods in the recent years. This has provided architectures for contextualized representation of an input text and generating fluent, diverseand human-like paraphrases. This paper surveys various approaches to paraphrase generation with a main focus on neural methods.

Exposure Bias versus Self-Recovery: Are Distortions Really Incremental for Autoregressive Text Generation?
Tianxing He et al.

Exposure bias has been regarded as a central problem for auto-regressive language models (LM). It claims that teacher forcing would cause the test-time generation to be incrementally distorted due to the training-generation discrepancy. Although a lot of algorithms have been proposed to avoid teacher forcing and therefore alleviate exposure bias, there is little work showing how serious the exposure bias problem actually is. In this work, we focus on the task of open-ended language generation, propose metrics to quantify the impact of exposure bias in the aspects of quality, diversity, and consistency. Our key intuition is that if we feed ground-truth data prefixes (instead of prefixes generated by the model itself) into the model and ask it to continue the generation, the performance should become much better because the training-generation discrepancy in the prefix is removed. Both automatic and human evaluations are conducted in our experiments. On the contrary to the popular belief in exposure bias, we find that the the distortion induced by the prefix discrepancy is limited, and does not seem to be incremental during the generation. Moreover, our analysis reveals an interesting self-recovery ability of the LM, which we hypothesize to be countering the harmful effects from exposure bias.

Generating Self-Contained and Summary-Centric Question Answer Pairs via Differentiable Reward Imitation Learning
Li Zhou et al.

Motivated by suggested question generation in conversational news recommendation systems, we propose a model for generating question-answer pairs (QA pairs) with self-contained, summary-centric questions and length-constrained, article-summarizing answers. We begin by collecting a new dataset of news articles with questions as titles and pairing them with summaries of varying length. This dataset is used to learn a QA pair generation model producing summaries as answers that balance brevity with sufficiency jointly with their cor-

responding questions. We then reinforce the QA pair generation process with a differentiable reward function to mitigate exposure bias, a common problem in natural language generation. Both automatic metrics and human evaluation demonstrate these QA pairs successfully capture the central gists of the articles and achieve high answer accuracy.

Unsupervised Paraphrasing with Pretrained Language Models
Tong Niu et al.

Paraphrase generation has benefited extensively from recent progress in the designing of training objectives and model architectures. However, previous explorations have largely focused on supervised methods, which require a large amount of labeled data that is costly to collect. To address this drawback, we adopt a transfer learning approach and propose a training pipeline that enables pre-trained language models to generate high-quality paraphrases in an unsupervised setting. Our recipe consists of task-adaptation, self-supervision, and a novel decoding algorithm named Dynamic Blocking (DB). To enforce a surface form dissimilar from the input, whenever the language model emits a token contained in the source sequence, DB prevents the model from outputting the subsequent source token for the next generation step. We show with automatic and human evaluations that our approach achieves state-of-the-art performance on both the Quora Question Pair (QQP) and the ParaNMT datasets and is robust to domain shift between the two datasets of distinct distributions. We also demonstrate that our model transfers to paraphrasing in other languages without any additional finetuning.

Profanity-Avoiding Training Framework for Seq2seq Models with Certified Robustness
Hengtong Zhang et al.

Seq2seq models have demonstrated their incredible effectiveness in a large variety of applications. However, recent research has shown that inappropriate language in training samples and well-designed testing cases can induce seq2seq models to output profanity. These outputs may potentially hurt the usability of seq2seq models and make the end-users feel offended. To address this problem, we propose a training framework with certified robustness to eliminate the causes that trigger the generation of profanity. The proposed training framework leverages merely a short list of profanity examples to prevent seq2seq models from generating a broader spectrum of profanity. The framework is composed of a pattern-eliminating training component to suppress the impact of language patterns with profanity in the training set, and a trigger-resisting training component to provide certified robustness for seq2seq models against intentionally injected profanity-triggering expressions in test samples. In the experiments, we consider two representative NLP tasks that seq2seq can be applied to, i.e., style transfer and dialogue generation. Extensive experimental results show that the proposed training framework can successfully prevent the NLP models from generating profanity.

Journalistic Guidelines Aware News Image Captioning
Xuewen Yang et al.

The task of news article image captioning aims to generate descriptive and informative captions for news article images. Unlike conventional image captions that simply describe the content of the image in general terms, news image captions follow journalistic guidelines and rely heavily on named entities to describe the image content, often drawing context from the whole article they are associated with. In this work, we propose a new approach to this task, motivated by caption guidelines that journalists follow. Our approach, Journalistic Guidelines Aware News Image Captioning (JoGANIC), leverages the structure of captions to improve the generation quality and guide our representation design. Experimental results, including detailed ablation studies, on two large-scale publicly available datasets show that JoGANIC substantially outperforms state-of-the-art methods both on caption generation and named entity related metrics.

AESOP: Paraphrase Generation with Adaptive Syntactic Control
Jiao Sun, Xuezhe Ma, and Nanyun Peng

We propose to control paraphrase generation through carefully chosen target syntactic structures to generate more proper and higher quality paraphrases. Our model, AESOP, leverages a pretrained language model and adds deliberately chosen syntactical control via a retrieval-based selection module to generate fluent paraphrases. Experiments show that AESOP achieves state-of-the-art performances on semantic preservation and syntactic conformation on two benchmark datasets with ground-truth syntactic control from human-annotated exemplars. Moreover, with the retrieval-based target syntax selection module, AESOP generates paraphrases with even better qualities than the current best model using human-annotated target syntactic parses according to human evaluation. We further demonstrate the effectiveness of AESOP to improve classification models' robustness to syntactic perturbation by data augmentation on two GLUE tasks.

Refocusing on Relevance: Personalization in NLG
Shiran Dudy, Steven Bedrick, and Bonnie Webber

Many NLG tasks such as summarization, dialogue response, or open domain question answering, focus primarily on a source text in order to generate a target response. This standard approach falls short, however, when a user's intent or context of work is not easily recoverable based solely on that source text– a scenario that we argue is more of the rule than the exception. In this work, we argue that NLG systems in general should place a much higher level of emphasis on making use of additional context, and suggest that relevance (as used in Information Retrieval) be thought of as a crucial tool for designing user-oriented text-generating tasks. We further discuss possible harms and hazards around such personalization, and argue that value-sensitive design represents a crucial path forward through these challenges.

The Future is not One-dimensional: Complex Event Schema Induction by Graph Modeling for Event Prediction
Manling Li et al.

Event schemas encode knowledge of stereotypical structures of events and their connections. As events unfold, schemas are crucial to act as a scaffolding. Previous work on event schema induction focuses either on atomic events or linear temporal event sequences, ignoring the interplay between events via arguments and argument relations. We introduce a new concept of Temporal Complex Event Schema: a graph-based schema representation that encompasses events, arguments, temporal connections and argument relations. In addition, we propose a Temporal Event Graph Model that predicts event instances following the temporal complex event schema. To build and evaluate such schemas, we release a new schema learning corpus containing 6,399 documents accompanied with event graphs, and we have manually constructed gold-standard schemas. Intrinsic evaluations by schema matching and instance graph perplexity, prove the superior quality of our probabilistic graph schema library compared to linear representations. Extrinsic evaluation on schema-guided future event prediction further demonstrates the predictive power of our event graph model, significantly outperforming human schemas and baselines by more than 17.8% on HITS1.

Learning Constraints and Descriptive Segmentation for Subevent Detection
Haoyu Wang et al.

Event mentions in text correspond to real-world events of varying degrees of granularity. The task of subevent detection aims to resolve this granularity issue, recognizing the membership of multi-granular events in event complexes. Since knowing the span of descriptive contexts of event complexes helps infer the membership of events, we propose the task of event-based text segmentation (EventSeg) as an auxiliary task to improve the learning for subevent detection. To bridge the two tasks together, we propose an approach to learning and enforcing constraints that capture dependencies between subevent detection and EventSeg prediction, as well as guiding the model to make globally consistent inference. Specifically, we adopt Rectifier Networks for constraint learning and then convert the learned constraints to a regularization term in the loss function of the neural model. Experimental results show that the proposed method outperforms baseline methods by 2.3% and 2.5% on benchmark datasets for subevent detection, HiEve and IC, respectively, while achieving a decent performance on EventSeg prediction.

ChemNER: Fine-Grained Chemistry Named Entity Recognition with Ontology-Guided Distant Supervision
Xuan Wang et al.

Scientific literature analysis needs fine-grained named entity recognition (NER) to provide a wide range of information for scientific discovery. For example, chemistry research needs to study dozens to hundreds of distinct, fine-grained entity types, making consistent and accurate annotation difficult even for crowds of domain experts. On the other hand, domain-specific ontologies and knowledge bases (KBs) can be easily accessed, constructed, or integrated, which makes distant supervision realistic for fine-grained chemistry NER. In distant supervision, training labels are generated by matching mentions in a document with the concepts in the knowledge bases (KBs). However, this kind of KB-matching suffers from two major challenges: incomplete annotation and noisy annotation. We propose ChemNER, an ontology-guided, distantly-supervised method for fine-grained chemistry NER to tackle these challenges. It leverages the chemistry type ontology structure to generate distant labels with novel methods of flexible KB-matching and ontology-guided multi-type disambiguation. It significantly improves the distant label generation for the subsequent sequence labeling model training. We also provide an expert-labeled, chemistry NER dataset with 62 fine-grained chemistry types (e.g., chemical compounds and chemical reactions). Experimental results show that ChemNER is highly effective, outperforming substantially the state-of-the-art NER methods (with .25 absolute F1 score improvement).

Moving on from OntoNotes: Coreference Resolution Model Transfer
Patrick Xia and Benjamin Van Durme

Academic neural models for coreference resolution (coref) are typically trained on a single dataset, OntoNotes, and model improvements are benchmarked on that same dataset. However, real-world applications of coref depend on the annotation guidelines and the domain of the target dataset, which often differ from those of OntoNotes. We aim to quantify transferability of coref models based on the number of annotated documents available in the target dataset. We examine eleven target datasets and find that continued training is consistently effective and especially beneficial when there are few target documents. We establish new benchmarks across several datasets, including state-of-the-art results on PreCo.

Document-level Entity-based Extraction as Template Generation
Kung-Hsiang Huang, Sam Tang, and Nanyun Peng

Document-level entity-based extraction (EE), aiming at extracting entity-centric information such as entity roles and entity relations, is key to automatic knowledge acquisition from text corpora for various domains. Most document-level EE systems build extractive models, which struggle to model long-term dependencies among entities at the document level. To address this issue, we propose a generative framework for two document-level EE tasks: role-filler entity extraction (REE) and relation extraction (RE). We first formulate them as a template generation problem, allowing models to efficiently capture cross-entity dependencies, exploit label semantics, and avoid the exponential computation complexity of identifying N-ary relations. A novel cross-attention guided copy mechanism, TopK Copy, is incorporated into a pre-trained sequence-to-

sequence model to enhance the capabilities of identifying key information in the input document. Experiments done on the MUC-4 and SciREX dataset show new state-of-the-art results on REE (+3.26%), binary RE (+4.8%), and 4-ary RE (+2.7%) in F1 score.

Lifelong Event Detection with Knowledge Transfer
Pengfei Yu, Heng Ji, and Prem Natarajan

Traditional supervised Information Extraction (IE) methods can extract structured knowledge elements from unstructured data, but it is limited to a pre-defined target ontology. In reality, the ontology of interest may change over time, adding emergent new types or more fine-grained subtypes. We propose a new lifelong learning framework to address this challenge. We focus on lifelong event detection as an exemplar case and propose a new problem formulation that is also generalizable to other IE tasks. In event detection and more general IE tasks, rich correlations or semantic relatedness exist among hierarchical knowledge element types. In our proposed framework, knowledge is being transferred between learned old event types and new event types. Specifically, we update old knowledge with new event types' mentions using a self-training loss. In addition, we aggregate old event types' representations based on their similarities with new event types to initialize the new event types' representations. Experimental results show that our framework outperforms competitive baselines with a 5.1% absolute gain in the F1 score. Moreover, our proposed framework can boost the F1 score for over 30% absolute gain on some new long-tail rare event types with few training instances. Our knowledge transfer module improves performance on both learned event types and new event types under the lifelong learning setting, showing that it helps consolidate old knowledge and improve novel knowledge acquisition.

Modular Self-Supervision for Document-Level Relation Extraction
Sheng Zhang et al.

Extracting relations across large text spans has been relatively underexplored in NLP, but it is particularly important for high-value domains such as biomedicine, where obtaining high recall of the latest findings is crucial for practical applications. Compared to conventional information extraction confined to short text spans, document-level relation extraction faces additional challenges in both inference and learning. Given longer text spans, state-of-the-art neural architectures are less effective and task-specific self-supervision such as distant supervision becomes very noisy. In this paper, we propose decomposing document-level relation extraction into relation detection and argument resolution, taking inspiration from Davidsonian semantics. This enables us to incorporate explicit discourse modeling and leverage modular self-supervision for each sub-problem, which is less noise-prone and can be further refined end-to-end via variational EM. We conduct a thorough evaluation in biomedical machine reading for precision oncology, where cross-paragraph relation mentions are prevalent. Our method outperforms prior state of the art, such as multi-scale learning and graph neural networks, by over 20 absolute F1 points. The gain is particularly pronounced among the most challenging relation instances whose arguments never co-occur in a paragraph.

Unsupervised Paraphrasing Consistency Training for Low Resource Named Entity Recognition
Rui Wang and Ricardo Henao

Unsupervised consistency training is a way of semi-supervised learning that encourages consistency in model predictions between the original and augmented data. For Named Entity Recognition (NER), existing approaches augment the input sequence with token replacement, assuming annotations on the replaced positions unchanged. In this paper, we explore the use of paraphrasing as a more principled data augmentation scheme for NER unsupervised consistency training. Specifically, we convert Conditional Random Field (CRF) into a multi-label classification module and encourage consistency on the entity appearance between the original and paraphrased sequences. Experiments show that our method is especially effective when annotations are limited.

Fine-grained Entity Typing without Knowledge Base
Jing Qian et al.

Existing work on Fine-grained Entity Typing (FET) typically trains automatic models on the datasets obtained by using Knowledge Bases (KB) as distant supervision. However, the reliance on KB means this training setting can be hampered by the lack of or the incompleteness of the KB. To alleviate this limitation, we propose a novel setting for training FET models: FET without accessing any knowledge base. Under this setting, we propose a two-step framework to train FET models. In the first step, we automatically create pseudo data with fine-grained labels from a large unlabeled dataset. Then a neural network model is trained based on the pseudo data, either in an unsupervised way or using self-training under the weak guidance from a coarse-grained Named Entity Recognition (NER) model. Experimental results show that our method achieves competitive performance with respect to the models trained on the original KB-supervised datasets.

Adversarial Attack against Cross-lingual Knowledge Graph Alignment
Zeru Zhang et al.

Recent literatures have shown that knowledge graph (KG) learning models are highly vulnerable to adversarial attacks. However, there is still a paucity of vulnerability analyses of cross-lingual entity alignment under adversarial attacks. This paper proposes an adversarial attack model with two novel attack techniques to perturb the KG structure and degrade the quality of deep cross-lingual entity alignment. First, an entity density

maximization method is employed to hide the attacked entities in dense regions in two KGs, such that the derived perturbations are unnoticeable. Second, an attack signal amplification method is developed to reduce the gradient vanishing issues in the process of adversarial attacks for further improving the attack effectiveness.

Towards Realistic Few-Shot Relation Extraction
Sam Brody, Sichao Wu, and Adrian Benton

In recent years, few-shot models have been applied successfully to a variety of NLP tasks. Han et al. (2018) introduced a few-shot learning framework for relation classification, and since then, several models have surpassed human performance on this task, leading to the impression that few-shot relation classification is solved. In this paper we take a deeper look at the efficacy of strong few-shot classification models in the more common relation extraction setting, and show that typical few-shot evaluation metrics obscure a wide variability in performance across relations. In particular, we find that state of the art few-shot relation classification models overly rely on entity type information, and propose modifications to the training routine to encourage models to better discriminate between relations involving similar entity types.

Data Augmentation for Cross-Domain Named Entity Recognition
Shuguang Chen et al.

Current work in named entity recognition (NER) shows that data augmentation techniques can produce more robust models. However, most existing techniques focus on augmenting in-domain data in low-resource scenarios where annotated data is quite limited. In this work, we take this research direction to the opposite and study cross-domain data augmentation for the NER task. We investigate the possibility of leveraging data from high-resource domains by projecting it into the low-resource domains. Specifically, we propose a novel neural architecture to transform the data representation from a high-resource to a low-resource domain by learning the patterns (e.g. style, noise, abbreviations, etc.) in the text that differentiate them and a shared feature space where both domains are aligned. We experiment with diverse datasets and show that transforming the data to the low-resource domain representation achieves significant improvements over only using data from high-resource domains.

Incorporating medical knowledge in BERT for clinical relation extraction
Arpita Roy and Shimei Pan

In recent years pre-trained language models (PLM) such as BERT have proven to be very effective in diverse NLP tasks such as Information Extraction, Sentiment Analysis and Question Answering. Trained with massive general-domain text, these pre-trained language models capture rich syntactic, semantic and discourse information in the text. However, due to the differences between general and specific domain text (e.g., Wikipedia versus clinic notes), these models may not be ideal for domain-specific tasks (e.g., extracting clinical relations). Furthermore, it may require additional medical knowledge to understand clinical text properly. To solve these issues, in this research, we conduct a comprehensive examination of different techniques to add medical knowledge into a pre-trained BERT model for clinical relation extraction. Our best model outperforms the state-of-the-art systems on the benchmark i2b2/VA 2010 clinical relation extraction dataset.

ECONET: Effective Continual Pretraining of Language Models for Event Temporal Reasoning
Rujun Han, Xiang Ren, and Nanyun Peng

While pre-trained language models (PTLMs) have achieved noticeable success on many NLP tasks, they still struggle for tasks that require event temporal reasoning, which is essential for event-centric applications. We present a continual pre-training approach that equips PTLMs with targeted knowledge about event temporal relations. We design self-supervised learning objectives to recover masked-out event and temporal indicators and to discriminate sentences from their corrupted counterparts (where event or temporal indicators got replaced). By further pre-training a PTLM with these objectives jointly, we reinforce its attention to event and temporal information, yielding enhanced capability on event temporal reasoning. This **E**ffective **CON**tinual pre-training framework for **E**vent **T**emporal reasoning (ECONET) improves the PTLMs' fine-tuning performances across five relation extraction and question answering tasks and achieves new or on-par state-of-the-art performances in most of our downstream tasks.

Learning from Noisy Labels for Entity-Centric Information Extraction
Wenxuan Zhou and Muhao Chen

Recent information extraction approaches have relied on training deep neural models. However, such models can easily overfit noisy labels and suffer from performance degradation. While it is very costly to filter noisy labels in large learning resources, recent studies show that such labels take more training steps to be memorized and are more frequently forgotten than clean labels, therefore are identifiable in training. Motivated by such properties, we propose a simple co-regularization framework for entity-centric information extraction, which consists of several neural models with identical structures but different parameter initialization. These models are jointly optimized with the task-specific losses and are regularized to generate similar predictions based on an agreement loss, which prevents overfitting on noisy labels. Extensive experiments on two widely used but noisy benchmarks for information extraction, TACRED and CoNLL03, demonstrate the effectiveness of our framework. We release our code to the community for future research.

Extracting Material Property Measurement Data from Scientific Articles
Gihan Panapitiya et al.

Machine learning-based prediction of material properties is often hampered by the lack of sufficiently large training data sets. The majority of such measurement data is embedded in scientific literature and the ability to automatically extract these data is essential to support the development of reliable property prediction methods. In this work, we describe a methodology for developing an automatic property extraction framework using material solubility as the target property. We create a training and evaluation data set containing tags for solubility-related entities using a combination of regular expressions and manual tagging. We then compare five entity recognition models leveraging both token-level and span-level architectures on the task of classifying solute names, solubility values, and solubility units. Additionally, we explore a novel pretraining approach that leverages automated chemical name and quantity extraction tools to generate large datasets that do not rely on intensive manual tagging. Finally, we perform an analysis to identify the causes of classification errors.

Modeling Document-Level Context for Event Detection via Important Context Selection
Amir Pouran Ben Veyseh et al.

The task of Event Detection (ED) in Information Extraction aims to recognize and classify trigger words of events in text. The recent progress has featured advanced transformer-based language models (e.g., BERT) as a critical component in state-of-the-art models for ED. However, the length limit for input texts is a barrier for such ED models as they cannot encode long-range document-level context that has been shown to be beneficial for ED. To address this issue, we propose a novel method to model document-level context for ED that dynamically selects relevant sentences in the document for the event prediction of the target sentence. The target sentence will be then augmented with the selected sentences and consumed entirely by transformer-based language models for improved representation learning for ED. To this end, the REINFORCE algorithm is employed to train the relevant sentence selection for ED. Several information types are then introduced to form the reward function for the training process, including ED performance, sentence similarity, and discourse relations. Our extensive experiments on multiple benchmark datasets reveal the effectiveness of the proposed model, leading to new state-of-the-art performance.

Crosslingual Transfer Learning for Relation and Event Extraction via Word Category and Class Alignments
Minh Van Nguyen et al.

Previous work on crosslingual Relation and Event Extraction (REE) suffers from the monolingual bias issue due to the training of models on only the source language data. An approach to overcome this issue is to use unlabeled data in the target language to aid the alignment of crosslingual representations, i.e., via fooling a language discriminator. However, as this approach does not condition on class information, a target language example of a class could be incorrectly aligned to a source language example of a different class. To address this issue, we propose a novel crosslingual alignment method that leverages class information of REE tasks for representation learning. In particular, we propose to learn two versions of representation vectors for each class in an REE task based on either source or target language examples. Representation vectors for corresponding classes will then be aligned to achieve class-aware alignment for crosslingual representations. In addition, we propose to further align representation vectors for language-universal word categories (i.e., parts of speech and dependency relations). As such, a novel filtering mechanism is presented to facilitate the learning of word category representations from contextualized representations on input texts based on adversarial learning. We conduct extensive crosslingual experiments with English, Chinese, and Arabic over REE tasks. The results demonstrate the benefits of the proposed method that significantly advances the state-of-the-art performance in these settings.

Corpus-based Open-Domain Event Type Induction
Jiaming Shen et al.

Traditional event extraction methods require predefined event types and their corresponding annotations to learn event extractors. These prerequisites are often hard to be satisfied in real-world applications. This work presents a corpus-based open-domain event type induction method that automatically discovers a set of event types from a given corpus. As events of the same type could be expressed in multiple ways, we propose to represent each event type as a cluster of <predicate sense, object head> pairs. Specifically, our method (1) selects salient predicates and object heads, (2) disambiguates predicate senses using only a verb sense dictionary, and (3) obtains event types by jointly embedding and clustering <predicate sense, object head> pairs in a latent spherical space. Our experiments, on three datasets from different domains, show our method can discover salient and high-quality event types, according to both automatic and human evaluations.

PDALN: Progressive Domain Adaptation over a Pre-trained Model for Low-Resource Cross-Domain Named Entity Recognition
Tao Zhang et al.

Cross-domain Named Entity Recognition (NER) transfers the NER knowledge from high-resource domains to the low-resource target domain. Due to limited labeled resources and domain shift, cross-domain NER is a challenging task. To address these challenges, we propose a progressive domain adaptation Knowledge Distillation (KD) approach – PDALN. It achieves superior domain adaptability by employing three components: (1) Adaptive data augmentation techniques, which alleviate cross-domain gap and label sparsity simultaneously; (2) Multi-level Domain invariant features, derived from a multi-grained MMD (Maximum Mean Discrepancy)

approach, to enable knowledge transfer across domains; (3) Advanced KD schema, which progressively enables powerful pre-trained language models to perform domain adaptation. Extensive experiments on four benchmarks show that PDALN can effectively adapt high-resource domains to low-resource target domains, even if they are diverse in terms and writing styles. Comparison with other baselines indicates the state-of-the-art performance of PDALN.

Multi-Vector Attention Models for Deep Re-ranking
Giulio Zhou and Jacob Devlin

Large-scale document retrieval systems often utilize two styles of neural network models which live at two different ends of the joint computation vs. accuracy spectrum. The first style is dual encoder (or two-tower) models, where the query and document representations are computed completely independently and combined with a simple dot product operation. The second style is cross-attention models, where the query and document features are concatenated in the input layer and all computation is based on the joint query-document representation. Dual encoder models are typically used for retrieval and deep re-ranking, while cross-attention models are typically used for shallow re-ranking. In this paper, we present a lightweight architecture that explores this joint cost vs. accuracy trade-off based on multi-vector attention (MVA). We thoroughly evaluate our method on the MS-MARCO passage retrieval dataset and show how to efficiently trade off retrieval accuracy with joint computation and offline document storage cost. We show that a highly compressed document representation and inexpensive joint computation can be achieved through a combination of learned pooling tokens and aggressive downprojection. Our code and model checkpoints are open-source and available on GitHub.

Toward Deconfounding the Influence of Subject's Demographic Characteristics in Question Answering
Maharshi Gor, Kellie Webster, and Jordan Boyd-Graber

The goal of question answering (QA) is to answer _any_ question. However, major QA datasets have skewed distributions over gender, profession, and nationality. Despite that skew, an analysis of model accuracy reveals little evidence that accuracy is lower for people based on gender or nationality; instead, there is more variation on professions (question topic) and question ambiguity. But QA's lack of representation could itself hide evidence of bias, necessitating QA datasets that better represent global diversity.

Exploring Strategies for Generalizable Commonsense Reasoning with Pre-trained Models
Kaixin Ma et al.

Commonsense reasoning benchmarks have been largely solved by fine-tuning language models. The downside is that fine-tuning may cause models to overfit to task-specific data and thereby forget their knowledge gained during pre-training. Recent works only propose lightweight model updates as models may already possess useful knowledge from past experience, but a challenge remains in understanding what parts and to what extent models should be refined for a given task. In this paper, we investigate what models learn from commonsense reasoning datasets. We measure the impact of three different adaptation methods on the generalization and accuracy of models. Our experiments with two models show that fine-tuning performs best, by learning both the content and the structure of the task, but suffers from overfitting and limited generalization to novel answers. We observe that alternative adaptation methods like prefix-tuning have comparable accuracy, but generalize better to unseen answers and are more robust to adversarial splits.

Transformer Feed-Forward Layers Are Key-Value Memories
Mor Geva et al.

Feed-forward layers constitute two-thirds of a transformer model's parameters, yet their role in the network remains under-explored. We show that feed-forward layers in transformer-based language models operate as key-value memories, where each key correlates with textual patterns in the training examples, and each value induces a distribution over the output vocabulary. Our experiments show that the learned patterns are human-interpretable, and that lower layers tend to capture shallow patterns, while upper layers learn more semantic ones. The values complement the keys' input patterns by inducing output distributions that concentrate probability mass on tokens likely to appear immediately after each pattern, particularly in the upper layers. Finally, we demonstrate that the output of a feed-forward layer is a composition of its memories, which is subsequently refined throughout the model's layers via residual connections to produce the final output distribution.

Connecting Attributions and QA Model Behavior on Realistic Counterfactuals
Xi Ye, Rohan Nair, and Greg Durrett

When a model attribution technique highlights a particular part of the input, a user might understand this highlight as making a statement about counterfactuals (Miller, 2019): if that part of the input were to change, the model's prediction might change as well. This paper investigates how well different attribution techniques align with this assumption on realistic counterfactuals in the case of reading comprehension (RC). RC is a particularly challenging test case, as token-level attributions that have been extensively studied in other NLP tasks such as sentiment analysis are less suitable to represent the reasoning that RC models perform. We construct counterfactual sets for three different RC settings, and through heuristics that can connect attribution methods' outputs to high-level model behavior, we can evaluate how useful different attribution methods and even different formats are for understanding counterfactuals. We find that pairwise attributions are better suited to RC than token-level attributions across these different RC settings, with our best performance coming from a modification that we propose to an existing pairwise attribution method.

How Do Neural Sequence Models Generalize? Local and Global Cues for Out-of-Distribution Prediction
D. Anthony Bau and Jacob Andreas

After a neural sequence model encounters an unexpected token, can its behavior be predicted? We show that RNN and transformer language models exhibit structured, consistent generalization in out-of-distribution contexts. We begin by introducing two idealized models of generalization in next-word prediction: a lexical context model in which generalization is consistent with the last word observed, and a syntactic context model in which generalization is consistent with the global structure of the input. In experiments in English, Finnish, Mandarin, and random regular languages, we demonstrate that neural language models interpolate between these two forms of generalization: their predictions are well-approximated by a log-linear combination of lexical and syntactic predictive distributions. We then show that, in some languages, noise mediates the two forms of generalization: noise applied to input tokens encourages syntactic generalization, while noise in history representations encourages lexical generalization. Finally, we offer a preliminary theoretical explanation of these results by proving that the observed interpolation behavior is expected in log-linear models with a particular feature correlation structure. These results help explain the effectiveness of two popular regularization schemes and show that aspects of sequence model generalization can be understood and controlled.

Comparing Text Representations: A Theory-Driven Approach
Gregory Yauney and David Mimno

Much of the progress in contemporary NLP has come from learning representations, such as masked language model (MLM) contextual embeddings, that turn challenging problems into simple classification tasks. But how do we quantify and explain this effect? We adapt general tools from computational learning theory to fit the specific characteristics of text datasets and present a method to evaluate the compatibility between representations and tasks. Even though many tasks can be easily solved with simple bag-of-words (BOW) representations, BOW does poorly on hard natural language inference tasks. For one such task we find that BOW cannot distinguish between real and randomized labelings, while pre-trained MLM representations show 72x greater distinction between real and random labelings than BOW. This method provides a calibrated, quantitative measure of the difficulty of a classification-based NLP task, enabling comparisons between representations without requiring empirical evaluations that may be sensitive to initializations and hyperparameters. The method provides a fresh perspective on the patterns in a dataset and the alignment of those patterns with specific labels.

Human Rationales as Attribution Priors for Explainable Stance Detection
Sahil Jayaram and Emily Allaway

As NLP systems become better at detecting opinions and beliefs from text, it is important to ensure not only that models are accurate but also that they arrive at their predictions in ways that align with human reasoning. In this work, we present a method for imparting human-like rationalization to a stance detection model using crowdsourced annotations on a small fraction of the training data. We show that in a data-scarce setting, our approach can improve the reasoning of a state-of-the-art classifier—particularly for inputs containing challenging phenomena such as sarcasm—at no cost in predictive performance. Furthermore, we demonstrate that attention weights surpass a leading attribution method in providing faithful explanations of our model's predictions, thus serving as a computationally cheap and reliable source of attributions for our model.

The Stem Cell Hypothesis: Dilemma behind Multi-Task Learning with Transformer Encoders
Han He and Jinho D. Choi

Multi-task learning with transformer encoders (MTL) has emerged as a powerful technique to improve performance on closely-related tasks for both accuracy and efficiency while a question still remains whether or not it would perform as well on tasks that are distinct in nature. We first present MTL results on five NLP tasks, POS, NER, DEP, CON, and SRL, and depict its deficiency over single-task learning. We then conduct an extensive pruning analysis to show that a certain set of attention heads get claimed by most tasks during MTL, who interfere with one another to fine-tune those heads for their own objectives. Based on this finding, we propose the Stem Cell Hypothesis to reveal the existence of attention heads naturally talented for many tasks that cannot be jointly trained to create adequate embeddings for all of those tasks. Finally, we design novel parameter-free probes to justify our hypothesis and demonstrate how attention heads are transformed across the five tasks during MTL through label analysis.

Text Counterfactuals via Latent Optimization and Shapley-Guided Search
Xiaoli Fern and Quintin Pope

We study the problem of generating counterfactual text for a classifier as a means for understanding and debugging classification. Given a textual input and a classification model, we aim to minimally alter the text to change the model's prediction. White-box approaches have been successfully applied to similar problems in vision where one can directly optimize the continuous input. Optimization-based approaches become difficult in the language domain due to the discrete nature of text. We bypass this issue by directly optimizing in the latent space and leveraging a language model to generate candidate modifications from optimized latent representations. We additionally use Shapley values to estimate the combinatoric effect of multiple changes. We then use these estimates to guide a beam search for the final counterfactual text. We achieve favorable performance compared to recent white-box and black-box baselines using human and automatic evaluations. Ablation studies show that both latent optimization and the use of Shapley values improve success rate and the

quality of the generated counterfactuals.

"Average" Approximates "First Principal Component"? An Empirical Analysis on Representations from Neural Language Models
Zihan Wang, Chengyu Dong, and Jingbo Shang

Contextualized representations based on neural language models have furthered the state of the art in various NLP tasks. Despite its great success, the nature of such representations remains a mystery. In this paper, we present an empirical property of these representations—"average" approximates "first principal component". Specifically, experiments show that the average of these representations shares almost the same direction as the first principal component of the matrix whose columns are these representations. We believe this explains why the average representation is always a simple yet strong baseline. Our further examinations show that this property also holds in more challenging scenarios, for example, when the representations are from a model right after its random initialization. Therefore, we conjecture that this property is intrinsic to the distribution of representations and not necessarily related to the input structure. We realize that these representations empirically follow a normal distribution for each dimension, and by assuming this is true, we demonstrate that the empirical property can be in fact derived mathematically.

Controlled Evaluation of Grammatical Knowledge in Mandarin Chinese Language Models
Yiwen Wang et al.

Prior work has shown that structural supervision helps English language models learn generalizations about syntactic phenomena such as subject-verb agreement. However, it remains unclear if such an inductive bias would also improve language models' ability to learn grammatical dependencies in typologically different languages. Here we investigate this question in Mandarin Chinese, which has a logographic, largely syllable-based writing system; different word order; and sparser morphology than English. We train LSTMs, Recurrent Neural Network Grammars, Transformer language models, and Transformer-parameterized generative parsing models on two Mandarin Chinese datasets of different sizes. We evaluate the models' ability to learn different aspects of Mandarin grammar that assess syntactic and semantic relationships. We find suggestive evidence that structural supervision helps with representing syntactic state across intervening content and improves performance in low-data settings, suggesting that the benefits of hierarchical inductive biases in acquiring dependency relationships may extend beyond English.

GradTS: A Gradient-Based Automatic Auxiliary Task Selection Method Based on Transformer Networks
Weicheng Ma et al.

A key problem in multi-task learning (MTL) research is how to select high-quality auxiliary tasks automatically. This paper presents GradTS, an automatic auxiliary task selection method based on gradient calculation in Transformer-based models. Compared to AUTOSEM, a strong baseline method, GradTS improves the performance of MT-DNN with a bert-base-cased backend model, from 0.33% to 17.93% on 8 natural language understanding (NLU) tasks in the GLUE benchmarks. GradTS is also time-saving since (1) its gradient calculations are based on single-task experiments and (2) the gradients are re-used without additional experiments when the candidate task set changes. On the 8 GLUE classification tasks, for example, GradTS costs on average 21.32% less time than AUTOSEM with comparable GPU consumption. Further, we show the robustness of GradTS across various task settings and model selections, e.g. mixed objectives among candidate tasks. The efficiency and efficacy of GradTS in these case studies illustrate its general applicability in MTL research without requiring manual task filtering or costly parameter tuning.

NegatER: Unsupervised Discovery of Negatives in Commonsense Knowledge Bases
Tara Safavi, Jing Zhu, and Danai Koutra

Codifying commonsense knowledge in machines is a longstanding goal of artificial intelligence. Recently, much progress toward this goal has been made with automatic knowledge base (KB) construction techniques. However, such techniques focus primarily on the acquisition of positive (true) KB statements, even though negative (false) statements are often also important for discriminative reasoning over commonsense KBs. As a first step toward the latter, this paper proposes NegatER, a framework that ranks potential negatives in commonsense KBs using a contextual language model (LM). Importantly, as most KBs do not contain negatives, NegatER relies only on the positive knowledge in the LM and does not require ground-truth negative examples. Experiments demonstrate that, compared to multiple contrastive data augmentation approaches, NegatER yields negatives that are more grammatical, coherent, and informative—leading to statistically significant accuracy improvements in a challenging KB completion task and confirming that the positive knowledge in LMs can be "re-purposed" to generate negative knowledge.

Instance-adaptive training with noise-robust losses against noisy labels
Lifeng Jin et al.

In order to alleviate the huge demand for annotated datasets for different tasks, many recent natural language processing datasets have adopted automated pipelines for fast-tracking usable data. However, model training with such datasets poses a challenge because popular optimization objectives are not robust to label noise induced in the annotation generation process. Several noise-robust losses have been proposed and evaluated on tasks in computer vision, but they generally use a single dataset-wise hyperparamter to control the strength of noise resistance. This work proposes novel instance-adaptive training frameworks to change single dataset-

wise hyperparameters of noise resistance in such losses to be instance-wise. Such instance-wise noise resistance hyperparameters are predicted by special instance-level label quality predictors, which are trained along with the main classification models. Experiments on noisy and corrupted NLP datasets show that proposed instance-adaptive training frameworks help increase the noise-robustness provided by such losses, promoting the use of the frameworks and associated losses in NLP models trained with noisy data.

Distributionally Robust Multilingual Machine Translation
Chunting Zhou et al.

Multilingual neural machine translation (MNMT) learns to translate multiple language pairs with a single model, potentially improving both the accuracy and the memory-efficiency of deployed models. However, the heavy data imbalance between languages hinders the model from performing uniformly across language pairs. In this paper, we propose a new learning objective for MNMT based on distributionally robust optimization, which minimizes the worst-case expected loss over the set of language pairs. We further show how to practically optimize this objective for large translation corpora using an iterated best response scheme, which is both effective and incurs negligible additional computational cost compared to standard empirical risk minimization. We perform extensive experiments on three sets of languages from two datasets and show that our method consistently outperforms strong baseline methods in terms of average and per-language performance under both many-to-one and one-to-many translation settings.

Model Selection for Cross-lingual Transfer
Yang Chen and Alan Ritter

Transformers that are pre-trained on multilingual corpora, such as, mBERT and XLM-RoBERTa, have achieved impressive cross-lingual transfer capabilities. In the zero-shot transfer setting, only English training data is used, and the fine-tuned model is evaluated on another target language. While this works surprisingly well, substantial variance has been observed in target language performance between different fine-tuning runs, and in the zero-shot setup, no target-language development data is available to select among multiple fine-tuned models. Prior work has relied on English dev data to select among models that are fine-tuned with different learning rates, number of steps and other hyperparameters, often resulting in suboptimal choices. In this paper, we show that it is possible to select consistently better models when small amounts of annotated data are available in auxiliary pivot languages. We propose a machine learning approach to model selection that uses the fine-tuned model's own internal representations to predict its cross-lingual capabilities. In extensive experiments we find that this method consistently selects better models than English validation data across twenty five languages (including eight low-resource languages), and often achieves results that are comparable to model selection using target language development data.

Continual Few-Shot Learning for Text Classification
Ramakanth Pasunuru, Veselin Stoyanov, and Mohit Bansal

Natural Language Processing (NLP) is increasingly relying on general end-to-end systems that need to handle many different linguistic phenomena and nuances. For example, a Natural Language Inference (NLI) system has to recognize sentiment, handle numbers, perform coreference, etc. Our solutions to complex problems are still far from perfect, so it is important to create systems that can learn to correct mistakes quickly, incrementally, and with little training data. In this work, we propose a continual few-shot learning (CFL) task, in which a system is challenged with a difficult phenomenon and asked to learn to correct mistakes with only a few (10 to 15) training examples. To this end, we first create benchmarks based on previously annotated data: two NLI (ANLI and SNLI) and one sentiment analysis (IMDB) datasets. Next, we present various baselines from diverse paradigms (e.g., memory-aware synapses and Prototypical networks) and compare them on few-shot learning and continual few-shot learning setups. Our contributions are in creating a benchmark suite and evaluation protocol for continual few-shot learning on the text classification tasks, and making several interesting observations on the behavior of similarity-based methods. We hope that our work serves as a useful starting point for future work on this important topic.

Efficient Nearest Neighbor Language Models
Junxian He, Graham Neubig, and Taylor Berg-Kirkpatrick

Non-parametric neural language models (NLMs) learn predictive distributions of text utilizing an external datastore, which allows them to learn through explicitly memorizing the training datapoints. While effective, these models often require retrieval from a large datastore at test time, significantly increasing the inference overhead and thus limiting the deployment of non-parametric NLMs in practical applications. In this paper, we take the recently proposed k-nearest neighbors language model as an example, exploring methods to improve its efficiency along various dimensions. Experiments on the standard WikiText-103 benchmark and domain-adaptation datasets show that our methods are able to achieve up to a 6x speed-up in inference speed while retaining comparable performance. The empirical analysis we present may provide guidelines for future research seeking to develop or deploy more efficient non-parametric NLMs.

STraTA: Self-Training with Task Augmentation for Better Few-shot Learning
Tu Vu et al.

Despite their recent successes in tackling many NLP tasks, large-scale pre-trained language models do not perform as well in few-shot settings where only a handful of training examples are available. To address this shortcoming, we propose STraTA, which stands for Self-Training with Task Augmentation, an approach that

builds on two key ideas for effective leverage of unlabeled data. First, STraTA uses task augmentation, a novel technique that synthesizes a large amount of data for auxiliary-task fine-tuning from target-task unlabeled texts. Second, STraTA performs self-training by further fine-tuning the strong base model created by task augmentation on a broad distribution of pseudo-labeled data. Our experiments demonstrate that STraTA can substantially improve sample efficiency across 12 few-shot benchmarks. Remarkably, on the SST-2 sentiment dataset, STraTA, with only 8 training examples per class, achieves comparable results to standard fine-tuning with 67K training examples. Our analyses reveal that task augmentation and self-training are both complementary and independently effective.

TADPOLE: Task ADapted Pre-Training via AnOmaLy DEtection
Vivek Madan, Ashish Khetan, and Zohar Karnin

The paradigm of pre-training followed by finetuning has become a standard procedure for NLP tasks, with a known problem of domain shift between the pre-training and downstream corpus. Previous works have tried to mitigate this problem with additional pre-training, either on the downstream corpus itself when it is large enough, or on a manually curated unlabeled corpus of a similar domain. In this paper, we address the problem for the case when the downstream corpus is too small for additional pre-training. We propose TADPOLE, a task adapted pre-training framework based on data selection techniques adapted from *Domain Adaptation*. We formulate the data selection as an anomaly detection problem that unlike existing methods works well when the downstream corpus is limited in size. It results in a scalable and efficient unsupervised technique that eliminates the need for any manual data curation. We evaluate our framework on eight tasks across four different domains: Biomedical, Computer Science, News, and Movie reviews, and compare its performance against competitive baseline techniques from the area of Domain Adaptation. Our framework outperforms all the baseline methods. On small datasets with less than 5K training examples, we get a gain of $1.82%$ in performance with additional pre-training for only $5%$ steps compared to the originally pre-trained models. It also compliments some of the other techniques such as data augmentation known for boosting performance when downstream corpus is small; highest performance is achieved when data augmentation is combined with task adapted pre-training.

Gradient-based Adversarial Attacks against Text Transformers
Chuan Guo et al.

We propose the first general-purpose gradient-based adversarial attack against transformer models. Instead of searching for a single adversarial example, we search for a distribution of adversarial examples parameterized by a continuous-valued matrix, hence enabling gradient-based optimization. We empirically demonstrate that our white-box attack attains state-of-the-art attack performance on a variety of natural language tasks, outperforming prior work in terms of adversarial success rate with matching imperceptibility as per automated and human evaluation. Furthermore, we show that a powerful black-box transfer attack, enabled by sampling from the adversarial distribution, matches or exceeds existing methods, while only requiring hard-label outputs.

Do Transformer Modifications Transfer Across Implementations and Applications?
Sharan Narang et al.

The research community has proposed copious modifications to the Transformer architecture since it was introduced over three years ago, relatively few of which have seen widespread adoption. In this paper, we comprehensively evaluate many of these modifications in a shared experimental setting that covers most of the common uses of the Transformer in natural language processing. Surprisingly, we find that most modifications do not meaningfully improve performance. Furthermore, most of the Transformer variants we found beneficial were either developed in the same codebase that we used or are relatively minor changes. We conjecture that performance improvements may strongly depend on implementation details and correspondingly make some recommendations for improving the generality of experimental results.

Paired Examples as Indirect Supervision in Latent Decision Models
Nitish Gupta et al.

Compositional, structured models are appealing because they explicitly decompose problems and provide interpretable intermediate outputs that give confidence that the model is not simply latching onto data artifacts. Learning these models is challenging, however, because end-task supervision only provides a weak indirect signal on what values the latent decisions should take. This often results in the model failing to learn to perform the intermediate tasks correctly. In this work, we introduce a way to leverage paired examples that provide stronger cues for learning latent decisions. When two related training examples share internal substructure, we add an additional training objective to encourage consistency between their latent decisions. Such an objective does not require external supervision for the values of the latent output, or even the end task, yet provides an additional training signal to that provided by individual training examples themselves. We apply our method to improve compositional question answering using neural module networks on the DROP dataset. We explore three ways to acquire paired questions in DROP: (a) discovering naturally occurring paired examples within the dataset, (b) constructing paired examples using templates, and (c) generating paired examples using a question generation model. We empirically demonstrate that our proposed approach improves both in- and out-of-distribution generalization and leads to correct latent decision predictions.

Pairwise Supervised Contrastive Learning of Sentence Representations
Dejiao Zhang et al.

Many recent successes in sentence representation learning have been achieved by simply fine-tuning on the Natural Language Inference (NLI) datasets with triplet loss or siamese loss. Nevertheless, they share a common weakness: sentences in a contradiction pair are not necessarily from different semantic categories. Therefore, optimizing the semantic entailment and contradiction reasoning objective alone is inadequate to capture the high-level semantic structure. The drawback is compounded by the fact that the vanilla siamese or triplet losses only learn from individual sentence pairs or triplets, which often suffer from bad local optima. In this paper, we propose PairSupCon, an instance discrimination based approach aiming to bridge semantic entailment and contradiction understanding with high-level categorical concept encoding. We evaluate PairSupCon on various downstream tasks that involve understanding sentence semantics at different granularities. We outperform the previous state-of-the-art method with 10%–13% averaged improvement on eight clustering tasks, and 5%–6% averaged improvement on seven semantic textual similarity (STS) tasks.

Muppet: Massive Multi-task Representations with Pre-Finetuning

Armen Aghajanyan et al.

We propose pre-finetuning, an additional large-scale learning stage between language model pre-training and fine-tuning. Pre-finetuning is massively multi-task learning (around 50 datasets, over 4.8 million total labeled examples), and is designed to encourage learning of representations that generalize better to many different tasks. We show that pre-finetuning consistently improves performance for pretrained discriminators (e.g.~RoBERTa) and generation models (e.g.~BART) on a wide range of tasks (sentence prediction, commonsense reasoning, MRC, etc.), while also significantly improving sample efficiency during fine-tuning. We also show that large-scale multi-tasking is crucial; pre-finetuning can hurt performance when few tasks are used up until a critical point (usually above 15) after which performance improves linearly in the number of tasks.

Diverse Distributions of Self-Supervised Tasks for Meta-Learning in NLP

Trapit Bansal et al.

Meta-learning considers the problem of learning an efficient learning process that can leverage its past experience to accurately solve new tasks. However, the efficacy of meta-learning crucially depends on the distribution of tasks available for training, and this is often assumed to be known a priori or constructed from limited supervised datasets. In this work, we aim to provide task distributions for meta-learning by considering self-supervised tasks automatically proposed from unlabeled text, to enable large-scale meta-learning in NLP. We design multiple distributions of self-supervised tasks by considering important aspects of task diversity, difficulty, type, domain, and curriculum, and investigate how they affect meta-learning performance. Our analysis shows that all these factors meaningfully alter the task distribution, some inducing significant improvements in downstream few-shot accuracy of the meta-learned models. Empirically, results on 20 downstream tasks show significant improvements in few-shot learning – adding up to +4.2% absolute accuracy (on average) to the previous unsupervised meta-learning method, and perform comparably to supervised methods on the FewRel 2.0 benchmark.

A Simple and Effective Method To Eliminate the Self Language Bias in Multilingual Representations

Ziyi Yang et al.

Language agnostic and semantic-language information isolation is an emerging research direction for multilingual representations models. We explore this problem from a novel angle of geometric algebra and semantic space. A simple but highly effective method "Language Information Removal (LIR)" factors out language identity information from semantic related components in multilingual representations pre-trained on multi-monolingual data. A post-training and model-agnostic method, LIR only uses simple linear operations, e.g. matrix factorization and orthogonal projection. LIR reveals that for weak-alignment multilingual systems, the principal components of semantic spaces primarily encodes language identity information. We first evaluate the LIR on a cross-lingual question answer retrieval task (LAReQA), which requires the strong alignment for the multilingual embedding space. Experiment shows that LIR is highly effectively on this task, yielding almost 100% relative improvement in MAP for weak-alignment models. We then evaluate the LIR on Amazon Reviews and XEVAL dataset, with the observation that removing language information is able to improve the cross-lingual transfer performance.

A Massively Multilingual Analysis of Cross-linguality in Shared Embedding Space

Alexander Jones, William Yang Wang, and Kyle Mahowald

In cross-lingual language models, representations for many different languages live in the same space. Here, we investigate the linguistic and non-linguistic factors affecting sentence-level alignment in cross-lingual pretrained language models for 101 languages and 5,050 language pairs. Using BERT-based LaBSE and BiLSTM-based LASER as our models, and the Bible as our corpus, we compute a task-based measure of cross-lingual alignment in the form of bitext retrieval performance, as well as four intrinsic measures of vector space alignment and isomorphism. We then examine a range of linguistic, quasi-linguistic, and training-related features as potential predictors of these alignment metrics. The results of our analyses show that word order agreement and agreement in morphological complexity are two of the strongest linguistic predictors of cross-linguality. We also note in-family training data as a stronger predictor than language-specific training data across the board. We verify some of our linguistic findings by looking at the effect of morphological segmentation on English-Inuktitut alignment, in addition to examining the effect of word order agreement on isomorphism for 66 zero-shot language pairs from a different corpus. We make the data and code for our experiments publicly available.

Frustratingly Simple but Surprisingly Strong: Using Language-Independent Features for Zero-shot Cross-lingual Semantic Parsing
Jingfeng Yang et al.

The availability of corpora has led to significant advances in training semantic parsers in English. Unfortunately, for languages other than English, annotated data is limited and so is the performance of the developed parsers. Recently, pretrained multilingual models have been proven useful for zero-shot cross-lingual transfer in many NLP tasks. What else does it require to apply a parser trained in English to other languages for zero-shot cross-lingual semantic parsing? Will simple language-independent features help? To this end, we experiment with six Discourse Representation Structure (DRS) semantic parsers in English, and generalize them to Italian, German and Dutch, where there are only a small number of manually annotated parses available. Extensive experiments show that despite its simplicity, adding Universal Dependency (UD) relations and Universal POS tags (UPOS) as model-agnostic features achieves surprisingly strong improvement on all parsers.

Improving Simultaneous Translation by Incorporating Pseudo-References with Fewer Reorderings
Junkun Chen et al.

Simultaneous translation is vastly different from full-sentence translation, in the sense that it starts translation before the source sentence ends, with only a few words delay. However, due to the lack of large-scale, high-quality simultaneous translation datasets, most such systems are still trained on conventional full-sentence bitexts. This is far from ideal for the simultaneous scenario due to the abundance of unnecessary long-distance reorderings in those bitexts. We propose a novel method that rewrites the target side of existing full-sentence corpora into simultaneous-style translation. Experiments on Zh→En and Ja→En simultaneous translation show substantial improvements (up to +2.7 BLEU) with the addition of these generated pseudo-references.

Classification-based Quality Estimation: Small and Efficient Models for Real-world Applications
Shuo Sun et al.

Sentence-level Quality estimation (QE) of machine translation is traditionally formulated as a regression task, and the performance of QE models is typically measured by Pearson correlation with human labels. Recent QE models have achieved previously-unseen levels of correlation with human judgments, but they rely on large multilingual contextualized language models that are computationally expensive and make them infeasible for real-world applications. In this work, we evaluate several model compression techniques for QE and find that, despite their popularity in other NLP tasks, they lead to poor performance in this regression setting. We observe that a full model parameterization is required to achieve SoTA results in a regression task. However, we argue that the level of expressiveness of a model in a continuous range is unnecessary given the downstream applications of QE, and show that reframing QE as a classification problem and evaluating QE models using classification metrics would better reflect their actual performance in real-world applications.

A Large-Scale Study of Machine Translation in Turkic Languages
Jamshidbek Mirzakhalov et al.

Recent advances in neural machine translation (NMT) have pushed the quality of machine translation systems to the point where they are becoming widely adopted to build competitive systems. However, there is still a large number of languages that are yet to reap the benefits of NMT. In this paper, we provide the first large-scale case study of the practical application of MT in the Turkic language family in order to realize the gains of NMT for Turkic languages under high-resource to extremely low-resource scenarios. In addition to presenting an extensive analysis that identifies the bottlenecks towards building competitive systems to ameliorate data scarcity, our study has several key contributions, including, i) a large parallel corpus covering 22 Turkic languages consisting of common public datasets in combination with new datasets of approximately 1.4 million parallel sentences, ii) bilingual baselines for 26 language pairs, iii) novel high-quality test sets in three different translation domains and iv) human evaluation scores. All models, scripts, and data will be released to the public.

Analyzing the Surprising Variability in Word Embedding Stability Across Languages
Laura Burdick, Jonathan K. Kummerfeld, and Rada Mihalcea

Word embeddings are powerful representations that form the foundation of many natural language processing architectures, both in English and in other languages. To gain further insight into word embeddings, we explore their stability (e.g., overlap between the nearest neighbors of a word in different embedding spaces) in diverse languages. We discuss linguistic properties that are related to stability, drawing out insights about correlations with affixing, language gender systems, and other features. This has implications for embedding use, particularly in research that uses them to study language trends.

Rule-based Morphological Inflection Improves Neural Terminology Translation
Weijia Xu and Marine Carpuat

Current approaches to incorporating terminology constraints in machine translation (MT) typically assume that the constraint terms are provided in their correct morphological forms. This limits their application to real-world scenarios where constraint terms are provided as lemmas. In this paper, we introduce a modular framework for incorporating lemma constraints in neural MT (NMT) in which linguistic knowledge and di-

verse types of NMT models can be flexibly applied. It is based on a novel cross-lingual inflection module that inflects the target lemma constraints based on the source context. We explore linguistically motivated rule-based and data-driven neural-based inflection modules and design English-German health and English-Lithuanian news test suites to evaluate them in domain adaptation and low-resource MT settings. Results show that our rule-based inflection module helps NMT models incorporate lemma constraints more accurately than a neural module and outperforms the existing end-to-end approach with lower training costs.

Good-Enough Example Extrapolation
Jason Wei

This paper asks whether extrapolating the hidden space distribution of text examples from one class onto another is a valid inductive bias for data augmentation. To operationalize this question, I propose a simple data augmentation protocol called "good-enough example extrapolation" (GE3). GE3 is lightweight and has no hyperparameters. Applied to three text classification datasets for various data imbalance scenarios, GE3 improves performance more than upsampling and other hidden-space data augmentation methods.

Learning to Selectively Learn for Weakly-supervised Paraphrase Generation
Kaize Ding et al.

Paraphrase generation is a longstanding NLP task that has diverse applications on downstream NLP tasks. However, the effectiveness of existing efforts predominantly relies on large amounts of golden labeled data. Though unsupervised endeavors have been proposed to alleviate this issue, they may fail to generate meaningful paraphrases due to the lack of supervision signals. In this work, we go beyond the existing paradigms and propose a novel approach to generate high-quality paraphrases with data of weak supervision. Specifically, we tackle the weakly-supervised paraphrase generation problem by: (1) obtaining abundant weakly-labeled parallel sentences via retrieval-based pseudo paraphrase expansion; and (2) developing a meta-learning framework to progressively select valuable samples for fine-tuning a pre-trained language model BART on the sentential paraphrasing task. We demonstrate that our approach achieves significant improvements over existing unsupervised approaches, and is even comparable in performance with supervised state-of-the-arts.

Effective Convolutional Attention Network for Multi-label Clinical Document Classification
Yang Liu et al.

Multi-label document classification (MLDC) problems can be challenging, especially for long documents with a large label set and a long-tail distribution over labels. In this paper, we present an effective convolutional attention network for the MLDC problem with a focus on medical code prediction from clinical documents. Our innovations are three-fold: (1) we utilize a deep convolution-based encoder with the squeeze-and-excitation networks and residual networks to aggregate the information across the document and learn meaningful document representations that cover different ranges of texts; (2) we explore multi-layer and sum-pooling attention to extract the most informative features from these multi-scale representations; (3) we combine binary cross entropy loss and focal loss to improve performance for rare labels. We focus our evaluation study on MIMIC-III, a widely used dataset in the medical domain. Our models outperform prior work on medical coding and achieve new state-of-the-art results on multiple metrics. We also demonstrate the language independent nature of our approach by applying it to two non-English datasets. Our model outperforms prior best model and a multilingual Transformer model by a substantial margin.

Contrastive Code Representation Learning
Paras Jain et al.

Recent work learns contextual representations of source code by reconstructing tokens from their context. For downstream semantic understanding tasks like code clone detection, these representations should ideally capture program functionality. However, we show that the popular reconstruction-based RoBERTa model is sensitive to source code edits, even when the edits preserve semantics. We propose ContraCode: a contrastive pre-training task that learns code functionality, not form. ContraCode pre-trains a neural network to identify functionally similar variants of a program among many non-equivalent distractors. We scalably generate these variants using an automated source-to-source compiler as a form of data augmentation. Contrastive pre-training outperforms RoBERTa on an adversarial code clone detection benchmark by 39% AUROC. Surprisingly, improved adversarial robustness translates to better accuracy over natural code; ContraCode improves summarization and TypeScript type inference accuracy by 2 to 13 percentage points over competitive baselines. All source is available at https://github.com/parasj/contracode.

IGA: An Intent-Guided Authoring Assistant
Simeng Sun et al.

While large-scale pretrained language models have significantly improved writing assistance functionalities such as autocomplete, more complex and controllable writing assistants have yet to be explored. We leverage advances in language modeling to build an interactive writing assistant that generates and rephrases text according to fine-grained author specifications. Users provide input to our Intent-Guided Assistant (IGA) in the form of text interspersed with tags that correspond to specific rhetorical directives (e.g., adding description or contrast, or rephrasing a particular sentence). We fine-tune a language model on a dataset heuristically-labeled with author intent, which allows IGA to fill in these tags with generated text that users can subsequently edit to their liking. A series of automatic and crowdsourced evaluations confirm the quality of IGA's generated

outputs, while a small-scale user study demonstrates author preference for IGA over baseline methods in a creative writing task. We release our dataset, code, and demo to spur further research into AI-assisted writing.

Math Word Problem Generation with Mathematical Consistency and Problem Context Constraints

Zichao Wang, Andrew Lan, and Richard Baraniuk

We study the problem of generating arithmetic math word problems (MWPs) given a math equation that specifies the mathematical computation and a context that specifies the problem scenario. Existing approaches are prone to generating MWPs that are either mathematically invalid or have unsatisfactory language quality. They also either ignore the context or require manual specification of a problem template, which compromises the diversity of the generated MWPs. In this paper, we develop a novel MWP generation approach that leverages i) pre-trained language models and a context keyword selection model to improve the language quality of generated MWPs and ii) an equation consistency constraint for math equations to improve the mathematical validity of the generated MWPs. Extensive quantitative and qualitative experiments on three real-world MWP datasets demonstrate the superior performance of our approach compared to various baselines.

Navigating the Kaleidoscope of COVID-19 Misinformation Using Deep Learning

Yuanzhi Chen and Mohammad Hasan

Irrespective of the success of the deep learning-based mixed-domain transfer learning approach for solving various Natural Language Processing tasks, it does not lend a generalizable solution for detecting misinformation from COVID-19 social media data. Due to the inherent complexity of this type of data, caused by its dynamic (context evolves rapidly), nuanced (misinformation types are often ambiguous), and diverse (skewed, fine-grained, and overlapping categories) nature, it is imperative for an effective model to capture both the local and global context of the target domain. By conducting a systematic investigation, we show that: (i) the deep Transformer-based pre-trained models, utilized via the mixed-domain transfer learning, are only good at capturing the local context, thus exhibits poor generalization, and (ii) a combination of shallow network-based domain-specific models and convolutional neural networks can efficiently extract local as well as global context directly from the target data in a hierarchical fashion, enabling it to offer a more generalizable solution.

Detecting Health Advice in Medical Research Literature

Yingya Li, Jun Wang, and Bei Yu

Health and medical researchers often give clinical and policy recommendations to inform health practice and public health policy. However, no current health information system supports the direct retrieval of health advice. This study fills the gap by developing and validating an NLP-based prediction model for identifying health advice in research publications. We annotated a corpus of 6,000 sentences extracted from structured abstracts in PubMed publications as ‘”strong advice”, “weak advice”, or “no advice”, and developed a BERT-based model that can predict, with a macro-averaged F1-score of 0.93, whether a sentence gives strong advice, weak advice, or not. The prediction model generalized well to sentences in both unstructured abstracts and discussion sections, where health advice normally appears. We also conducted a case study that applied this prediction model to retrieve specific health advice on COVID-19 treatments from LitCovid, a large COVID research literature portal, demonstrating the usefulness of retrieving health advice sentences as an advanced research literature navigation function for health researchers and the general public.

A Semantic Feature-Wise Transformation Relation Network for Automatic Short Answer Grading

Zhaohui Li, Yajur Tomar, and Rebecca J. Passonneau

Automatic short answer grading (ASAG) is the task of assessing students’ short natural language responses to objective questions. It is a crucial component of new education platforms, and could support more wide-spread use of constructed response questions to replace cognitively less challenging multiple choice questions. We propose a Semantic Feature-wise transformation Relation Network (SFRN) that exploits the multiple components of ASAG datasets more effectively. SFRN captures relational knowledge among the questions (Q), reference answers or rubrics (R), and labeled student answers (A). A relation network learns vector representations for the elements of QRA triples, then combines the learned representations using learned semantic feature-wise transformations. We apply translation-based data augmentation to address the two problems of limited training data, and high data skew for multi-class ASAG tasks. Our model has up to 11% performance improvement over state-of-the-art results on the benchmark SemEval-2013 datasets, and surpasses custom approaches designed for a Kaggle challenge, demonstrating its generality.

Evaluating Scholarly Impact: Towards Content-Aware Bibliometrics

Saurav Manchanda and George Karypis

Quantitatively measuring the impact-related aspects of scientific, engineering, and technological (SET) innovations is a fundamental problem with broad applications. Traditional citation-based measures for assessing the impact of innovations and related entities do not take into account the content of the publications. This limits their ability to provide rigorous quality-related metrics because they cannot account for the reasons that led to a citation. We present approaches to estimate content-aware bibliometrics to quantitatively measure the scholarly impact of a publication. Our approaches assess the impact of a cited publication by the extent to which the cited publication informs the citing publication. We introduce a new metric, called “Content Informed Index” (CII), that uses the content of the paper as a source of distant-supervision, to quantify how

much the cited-node informs the citing-node. We evaluate the weights estimated by our approach on three manually annotated datasets, where the annotations quantify the extent of information in the citation. Particularly, we evaluate how well the ranking imposed by our approach associates with the ranking imposed by the manual annotations. CII achieves up to 103% improvement in performance as compared to the second-best performing approach.

A Scalable Framework for Learning From Implicit User Feedback to Improve Natural Language Understanding in Large-Scale Conversational AI Systems
Sunghyun Park et al.

Natural Language Understanding (NLU) is an established component within a conversational AI or digital assistant system, and it is responsible for producing semantic understanding of a user request. We propose a scalable and automatic approach for improving NLU in a large-scale conversational AI system by leveraging implicit user feedback, with an insight that user interaction data and dialog context have rich information embedded from which user satisfaction and intention can be inferred. In particular, we propose a domain-agnostic framework for curating new supervision data for improving NLU from live production traffic. With an extensive set of experiments, we show the results of applying the framework and improving NLU for a large-scale production system across 10 domains.

Summarize-then-Answer: Generating Concise Explanations for Multi-hop Reading Comprehension
Naoya Inoue et al.

How can we generate concise explanations for multi-hop Reading Comprehension (RC)? The current strategies of identifying supporting sentences can be seen as an extractive question-focused summarization of the input text. However, these extractive explanations are not necessarily concise i.e. not minimally sufficient for answering a question. Instead, we advocate for an abstractive approach, where we propose to generate a question-focused, abstractive summary of input paragraphs and then feed it to an RC system. Given a limited amount of human-annotated abstractive explanations, we train the abstractive explainer in a semi-supervised manner, where we start from the supervised model and then train it further through trial and error maximizing a conciseness-promoted reward function. Our experiments demonstrate that the proposed abstractive explainer can generate more compact explanations than an extractive explainer with limited supervision (only 2k instances) while maintaining sufficiency.

FewshotQA: A simple framework for few-shot learning of question answering tasks using pre-trained text-to-text models
Rakesh Chada and Pradeep Natarajan

The task of learning from only a few examples (called a few-shot setting) is of key importance and relevance to a real-world setting. For question answering (QA), the current state-of-the-art pre-trained models typically need fine-tuning on tens of thousands of examples to obtain good results. Their performance degrades significantly in a few-shot setting (< 100 examples). To address this, we propose a simple fine-tuning framework that leverages pre-trained text-to-text models and is directly aligned with their pre-training framework. Specifically, we construct the input as a concatenation of the question, a mask token representing the answer span and a context. Given this input, the model is fine-tuned using the same objective as that of its pre-training objective. Through experimental studies on various few-shot configurations, we show that this formulation leads to significant gains on multiple QA benchmarks (an absolute gain of 34.2 F1 points on average when there are only 16 training examples). The gains extend further when used with larger models (Eg:- 72.3 F1 on SQuAD using BART-large with only 32 examples) and translate well to a multilingual setting . On the multilingual TydiQA benchmark, our model outperforms the XLM-Roberta-large by an absolute margin of upto 40 F1 points and an average of 33 F1 points in a few-shot setting (<= 64 training examples). We conduct detailed ablation studies to analyze factors contributing to these gains.

Multi-stage Training with Improved Negative Contrast for Neural Passage Retrieval
Jing Lu et al.

In the context of neural passage retrieval, we study three promising techniques: synthetic data generation, negative sampling, and fusion. We systematically investigate how these techniques contribute to the performance of the retrieval system and how they complement each other. We propose a multi-stage framework comprising of pre-training with synthetic data, fine-tuning with labeled data, and negative sampling at both stages. We study six negative sampling strategies and apply them to the fine-tuning stage and, as a noteworthy novelty, to the synthetic data that we use for pre-training. Also, we explore fusion methods that combine negatives from different strategies. We evaluate our system using two passage retrieval tasks for open-domain QA and using MS MARCO. Our experiments show that augmenting the negative contrast in both stages is effective to improve passage retrieval accuracy and, importantly, they also show that synthetic data generation and negative sampling have additive benefits. Moreover, using the fusion of different kinds allows us to reach performance that establishes a new state-of-the-art level in two of the tasks we evaluated.

Perhaps PTLMs Should Go to School — A Task to Assess Open Book and Closed Book QA
Manuel Ciosici et al.

Our goal is to deliver a new task and leaderboard to stimulate research on question answering and pre-trained language models (PTLMs) to understand a significant instructional document, e.g., an introductory college

textbook or a manual. PTLMs have shown great success in many question-answering tasks, given significant supervised training, but much less so in zero-shot settings. We propose a new task that includes two college-level introductory texts in the social sciences (American Government 2e) and humanities (U.S. History), hundreds of true/false statements based on review questions written by the textbook authors, validation/development tests based on the first eight chapters of the textbooks, blind tests based on the remaining textbook chapters, and baseline results given state-of-the-art PTLMs. Since the questions are balanced, random performance should be ~50%. T5, fine-tuned with BoolQ achieves the same performance, suggesting that the textbook's content is not pre-represented in the PTLM. Taking the exam closed book, but having read the textbook (i.e., adding the textbook to T5's pre-training), yields at best minor improvement (56%), suggesting that the PTLM may not have "understood" the textbook (or perhaps misunderstood the questions). Performance is better (~60%) when the exam is taken open-book (i.e., allowing the machine to automatically retrieve a paragraph and use it to answer the question).

ReasonBERT: Pre-trained to Reason with Distant Supervision
Xiang Deng et al.

We present ReasonBert, a pre-training method that augments language models with the ability to reason over long-range relations and multiple, possibly hybrid contexts. Unlike existing pre-training methods that only harvest learning signals from local contexts of naturally occurring texts, we propose a generalized notion of distant supervision to automatically connect multiple pieces of text and tables to create pre-training examples that require long-range reasoning. Different types of reasoning are simulated, including intersecting multiple pieces of evidence, bridging from one piece of evidence to another, and detecting unanswerable cases. We conduct a comprehensive evaluation on a variety of extractive question answering datasets ranging from single-hop to multi-hop and from text-only to table-only to hybrid that require various reasoning capabilities and show that ReasonBert achieves remarkable improvement over an array of strong baselines. Few-shot experiments further demonstrate that our pre-training method substantially improves sample efficiency.

Single-dataset Experts for Multi-dataset Question Answering
Dan Friedman, Ben Dodge, and Danqi Chen

Many datasets have been created for training reading comprehension models, and a natural question is whether we can combine them to build models that (1) perform better on all of the training datasets and (2) generalize and transfer better to new datasets. Prior work has addressed this goal by training one network simultaneously on multiple datasets, which works well on average but is prone to over- or under-fitting different sub- distributions and might transfer worse compared to source models with more overlap with the target dataset. Our approach is to model multi-dataset question answering with an ensemble of single-dataset experts, by training a collection of lightweight, dataset-specific adapter modules (Houlsby et al., 2019) that share an underlying Transformer model. We find that these Multi-Adapter Dataset Experts (MADE) outperform all our baselines in terms of in-distribution accuracy, and simple methods based on parameter-averaging lead to better zero-shot generalization and few-shot transfer performance, offering a strong and versatile starting point for building new reading comprehension systems.

Simple Entity-Centric Questions Challenge Dense Retrievers
Christopher Sciavolino et al.

Open-domain question answering has exploded in popularity recently due to the success of dense retrieval models, which have surpassed sparse models using only a few supervised training examples. However, in this paper, we demonstrate current dense models are not yet the holy grail of retrieval. We first construct EntityQuestions, a set of simple, entity-rich questions based on facts from Wikidata (e.g., "Where was Arve Furset born?"), and observe that dense retrievers drastically under-perform sparse methods. We investigate this issue and uncover that dense retrievers can only generalize to common entities unless the question pattern is explicitly observed during training. We discuss two simple solutions towards addressing this critical problem. First, we demonstrate that data augmentation is unable to fix the generalization problem. Second, we argue a more robust passage encoder helps facilitate better question adaptation using specialized question encoders. We hope our work can shed light on the challenges in creating a robust, universal dense retriever that works well across different input distributions.

Mitigating False-Negative Contexts in Multi-document Question Answering with Retrieval Marginalization
Ansong Ni, Matt Gardner, and Pradeep Dasigi

Question Answering (QA) tasks requiring information from multiple documents often rely on a retrieval model to identify relevant information for reasoning. The retrieval model is typically trained to maximize the likelihood of the labeled supporting evidence. However, when retrieving from large text corpora such as Wikipedia, the correct answer can often be obtained from multiple evidence candidates. Moreover, not all such candidates are labeled as positive during annotation, rendering the training signal weak and noisy. This problem is exacerbated when the questions are unanswerable or when the answers are Boolean, since the model cannot rely on lexical overlap to make a connection between the answer and supporting evidence. We develop a new parameterization of set-valued retrieval that handles unanswerable queries, and we show that marginalizing over this set during training allows a model to mitigate false negatives in supporting evidence annotations. We test our method on two multi-document QA datasets, IIRC and HotpotQA. On IIRC, we show that joint modeling with marginalization improves model performance by 5.5 F1 points and achieves a new state-of-the-art

performance of 50.5 F1. We also show that retrieval marginalization results in 4.1 QA F1 improvement over a non-marginalized baseline on HotpotQA in the fullwiki setting.

MultiDoc2Dial: Modeling Dialogues Grounded in Multiple Documents
Song Feng et al.

We propose MultiDoc2Dial, a new task and dataset on modeling goal-oriented dialogues grounded in multiple documents. Most previous works treat document-grounded dialogue modeling as machine reading comprehension task based on a single given document or passage. In this work, we aim to address more realistic scenarios where a goal-oriented information-seeking conversation involves multiple topics, and hence is grounded on different documents. To facilitate such task, we introduce a new dataset that contains dialogues grounded in multiple documents from four different domains. We also explore modeling the dialogue-based and document-based contexts in the dataset. We present strong baseline approaches and various experimental results, aiming to support further research efforts on such a task.

GupShup: Summarizing Open-Domain Code-Switched Conversations
Laiba Mehnaz et al.

Code-switching is the communication phenomenon where the speakers switch between different languages during a conversation. With the widespread adoption of conversational agents and chat platforms, code-switching has become an integral part of written conversations in many multi-lingual communities worldwide. Therefore, it is essential to develop techniques for understanding and summarizing these conversations. Towards this objective, we introduce the task of abstractive summarization of Hindi-English (Hi-En) code-switched conversations. We also develop the first code-switched conversation summarization dataset - *GupShup*, which contains over 6,800 Hi-En conversations and their corresponding human-annotated summaries in English (En) and Hi-En. We present a detailed account of the entire data collection and annotation process. We analyze the dataset using various code-switching statistics. We train state-of-the-art abstractive summarization models and report their performances using both automated metrics and human evaluation. Our results show that multi-lingual mBART and multi-view seq2seq models obtain the best performances on this new dataset. We also conduct an extensive qualitative analysis to provide insight into the models and some of their shortcomings.

BiSECT: Learning to Split and Rephrase Sentences with Bitexts
Joongwon Kim et al.

An important task in NLP applications such as sentence simplification is the ability to take a long, complex sentence and split it into shorter sentences, rephrasing as necessary. We introduce a novel dataset and a new model for this 'split and rephrase' task. Our BiSECT training data consists of 1 million long English sentences paired with shorter, meaning-equivalent English sentences. We obtain these by extracting 1-2 sentence alignments in bilingual parallel corpora and then using machine translation to convert both sides of the corpus into the same language. BiSECT contains higher quality training examples than the previous Split and Rephrase corpora, with sentence splits that require more significant modifications. We categorize examples in our corpus and use these categories in a novel model that allows us to target specific regions of the input sentence to be split and edited. Moreover, we show that models trained on BiSECT can perform a wider variety of split operations and improve upon previous state-of-the-art approaches in automatic and human evaluations.

Data Collection vs. Knowledge Graph Completion: What is Needed to Improve Coverage?
Kenneth Church and Yuchen Bian

This survey/position paper discusses ways to improve coverage of resources such as WordNet. Rapp estimated correlations, rho, between corpus statistics and pyscholinguistic norms. rho improves with quantity (corpus size) and quality (balance). 1M words is enough for simple estimates (unigram frequencies), but at least 100x more is required for good estimates of word associations and embeddings. Given such estimates, WordNet's coverage is remarkable. WordNet was developed on SemCor, a small sample (200k words) from the Brown Corpus. Knowledge Graph Completion (KGC) attempts to learn missing links from subsets. But Rapp's estimates of sizes suggest it would be more profitable to collect more data than to infer missing information that is not there.

Universal Sentence Representation Learning with Conditional Masked Language Model
Ziyi Yang et al.

This paper presents a novel training method, Conditional Masked Language Modeling (CMLM), to effectively learn sentence representations on large scale unlabeled corpora. CMLM integrates sentence representation learning into MLM training by conditioning on the encoded vectors of adjacent sentences. Our English CMLM model achieves state-of-the-art performance on SentEval, even outperforming models learned using supervised signals. As a fully unsupervised learning method, CMLM can be conveniently extended to a broad range of languages and domains. We find that a multilingual CMLM model co-trained with bitext retrieval (BR) and natural language inference (NLI) tasks outperforms the previous state-of-the-art multilingual models by a large margin, e.g. 10% improvement upon baseline models on cross-lingual semantic search. We explore the same language bias of the learned representations, and propose a simple, post-training and model agnostic approach to remove the language identifying information from the representation while still retaining sentence semantics.

On the Benefit of Syntactic Supervision for Cross-lingual Transfer in Semantic Role Labeling
Zhisong Zhang, Emma Strubell, and Eduard Hovy

Although recent developments in neural architectures and pre-trained representations have greatly increased state-of-the-art model performance on fully-supervised semantic role labeling (SRL), the task remains challenging for languages where supervised SRL training data are not abundant. Cross-lingual learning can improve performance in this setting by transferring knowledge from high-resource languages to low-resource ones. Moreover, we hypothesize that annotations of syntactic dependencies can be leveraged to further facilitate cross-lingual transfer. In this work, we perform an empirical exploration of the helpfulness of syntactic supervision for crosslingual SRL within a simple multitask learning scheme. With comprehensive evaluations across ten languages (in addition to English) and three SRL benchmark datasets, including both dependency- and span-based SRL, we show the effectiveness of syntactic supervision in low-resource scenarios.

Implicit Premise Generation with Discourse-aware Commonsense Knowledge Models
Tuhin Chakrabarty, Aadit Trivedi, and Smaranda Muresan

Enthymemes are defined as arguments where a premise or conclusion is left implicit. We tackle the task of generating the *implicit premise in an enthymeme*, which requires not only an understanding of the stated conclusion and premise but also additional inferences that could depend on commonsense knowledge. The largest available dataset for enthymemes (Habernal et al., 2018) consists of 1.7k samples, which is not large enough to train a neural text generation model. To address this issue, we take advantage of a similar task and dataset: Abductive reasoning in narrative text (Bhagavatula et al., 2020). However, we show that simply using a state-of-the-art seq2seq model fine-tuned on this data might not generate meaningful implicit premises associated with the given enthymemes. We demonstrate that encoding discourse-aware commonsense during fine-tuning improves the quality of the generated implicit premises and outperforms all other baselines both in automatic and human evaluations on three different datasets.

Inducing Transformer's Compositional Generalization Ability via Auxiliary Sequence Prediction Tasks
Yichen Jiang and Mohit Bansal

Systematic compositionality is an essential mechanism in human language, allowing the recombination of known parts to create novel expressions. However, existing neural models have been shown to lack this basic ability in learning symbolic structures. Motivated by the failure of a Transformer model on the SCAN compositionality challenge (Lake and Baroni, 2018), which requires parsing a command into actions, we propose two auxiliary sequence prediction tasks as additional training supervision. These automatically-generated sequences are more representative of the underlying compositional symbolic structures of the input data. During inference, the model jointly predicts the next action and the next tokens in the auxiliary sequences at each step. Experiments on the SCAN dataset show that our method encourages the Transformer to understand compositional structures of the command, improving its accuracy on multiple challenging splits from 10% to 100%. With only 418 (5%) training instances, our approach still achieves 97.8% accuracy on the MCD1 split. Therefore, we argue that compositionality can be induced in Transformers given minimal but proper guidance. We also show that a better result is achieved using less contextualized vectors as the attention's query, providing insights into architecture choices in achieving systematic compositionality. Finally, we show positive generalization results on the grounded-SCAN task (Ruis et al., 2020).

Flexible Generation of Natural Language Deductions
Kaj Bostrom et al.

An interpretable system for open-domain reasoning needs to express its reasoning process in a transparent form. Natural language is an attractive representation for this purpose — it is both highly expressive and easy for humans to understand. However, manipulating natural language statements in logically consistent ways is hard: models must cope with variation in how meaning is expressed while remaining precise. In this paper, we describe ParaPattern, a method for building models to generate deductive inferences from diverse natural language inputs without direct human supervision. We train BART-based models (Lewis et al., 2020) to generate the result of applying a particular logical operation to one or more premise statements. Crucially, we develop a largely automated pipeline for constructing suitable training examples from Wikipedia. We evaluate our models using out-of-domain sentence compositions from the QASC (Khot et al., 2020) and EntailmentBank (Dalvi et al., 2021) datasets as well as targeted perturbation sets. Our results show that our models are substantially more accurate and flexible than baseline systems. ParaPattern achieves 85% validity on examples of the 'substitution' operation from EntailmentBank without the use of any in-domain training data, matching the performance of a model fine-tuned for EntailmentBank. The full source code for our method is publicly available.

Structure-aware Fine-tuning of Sequence-to-sequence Transformers for Transition-based AMR Parsing
Jiawei Zhou et al.

Predicting linearized Abstract Meaning Representation (AMR) graphs using pre-trained sequence-to-sequence Transformer models has recently led to large improvements on AMR parsing benchmarks. These parsers are simple and avoid explicit modeling of structure but lack desirable properties such as graph well-formedness guarantees or built-in graph-sentence alignments. In this work we explore the integration of general pre-trained sequence-to-sequence language models and a structure-aware transition-based approach. We depart from a

pointer-based transition system and propose a simplified transition set, designed to better exploit pre-trained language models for structured fine-tuning. We also explore modeling the parser state within the pre-trained encoder-decoder architecture and different vocabulary strategies for the same purpose. We provide a detailed comparison with recent progress in AMR parsing and show that the proposed parser retains the desirable properties of previous transition-based approaches, while being simpler and reaching the new parsing state of the art for AMR 2.0, without the need for graph re-categorization.

Think about it! Improving defeasible reasoning by first modeling the question scenario.
Aman Madaan et al.

Defeasible reasoning is the mode of reasoning where conclusions can be overturned by taking into account new evidence. Existing cognitive science literature on defeasible reasoning suggests that a person forms a "mental model" of the problem scenario before answering questions. Our research goal asks whether neural models can similarly benefit from envisioning the question scenario before answering a defeasible query. Our approach is, given a question, to have a model first create a graph of relevant influences, and then leverage that graph as an additional input when answering the question. Our system, CURIOUS, achieves a new state-of-the-art on three different defeasible reasoning datasets. This result is significant as it illustrates that performance can be improved by guiding a system to "think about" a question and explicitly model the scenario, rather than answering reflexively.

Open Aspect Target Sentiment Classification with Natural Language Prompts
Ronald Seoh et al.

For many business applications, we often seek to analyze sentiments associated with any arbitrary aspects of commercial products, despite having a very limited amount of labels or even without any labels at all. However, existing aspect target sentiment classification (ATSC) models are not trainable if annotated datasets are not available. Even with labeled data, they fall short of reaching satisfactory performance. To address this, we propose simple approaches that better solve ATSC with natural language prompts, enabling the task under zero-shot cases and enhancing supervised settings, especially for few-shot cases. Under the few-shot setting for SemEval 2014 Task 4 laptop domain, our method of reformulating ATSC as an NLI task outperforms supervised SOTA approaches by up to 24.13 accuracy points and 33.14 macro F1 points. Moreover, we demonstrate that our prompts could handle implicitly stated aspects as well: our models reach about 77% accuracy on detecting sentiments for aspect categories (e.g., food), which do not necessarily appear within the text, even though we trained the models only with explicitly mentioned aspect terms (e.g., fajitas) from just 16 reviews - while the accuracy of the no-prompt baseline is only around 65%.

Does BERT Learn as Humans Perceive? Understanding Linguistic Styles through Lexica
Shirley Hayati, Dongyeop Kang, and Lyle Ungar

People convey their intention and attitude through linguistic styles of the text that they write. In this study, we investigate lexicon usages across styles throughout two lenses: human perception and machine word importance, since words differ in the strength of the stylistic cues that they provide. To collect labels of human perception, we curate a new dataset, Hummingbird, on top of benchmarking style datasets. We have crowd workers highlight the representative words in the text that makes them think the text has the following styles: politeness, sentiment, offensiveness, and five emotion types. We then compare these human word labels with word importance derived from a popular fine-tuned style classifier like BERT. Our results show that the BERT often finds content words not relevant to the target style as important words used in style prediction, but humans do not perceive the same way even though for some styles (e.g., positive sentiment and joy) human- and machine-identified words share significant overlap for some styles.

Improving Stance Detection with Multi-Dataset Learning and Knowledge Distillation
Yingjie Li, Chenye Zhao, and Cornelia Caragea

Stance detection determines whether the author of a text is in favor of, against or neutral to a specific target and provides valuable insights into important events such as legalization of abortion. Despite significant progress on this task, one of the remaining challenges is the scarcity of annotations. Besides, most previous works focused on a hard-label training in which meaningful similarities among categories are discarded during training. To address these challenges, first, we evaluate a multi-target and a multi-dataset training settings by training one model on each dataset and datasets of different domains, respectively. We show that models can learn more universal representations with respect to targets in these settings. Second, we investigate the knowledge distillation in stance detection and observe that transferring knowledge from a teacher model to a student model can be beneficial in our proposed training settings. Moreover, we propose an Adaptive Knowledge Distillation (AKD) method that applies instance-specific temperature scaling to the teacher and student predictions. Results show that the multi-dataset model performs best on all datasets and it can be further improved by the proposed AKD, outperforming the state-of-the-art by a large margin. We publicly release our code.

Discovering the Unknown Knowns: Turning Implicit Knowledge in the Dataset into Explicit Training Examples for Visual Question Answering
Jihyung Kil et al.

Visual question answering (VQA) is challenging not only because the model has to handle multi-modal information, but also because it is just so hard to collect sufficient training examples — there are too many questions one can ask about an image. As a result, a VQA model trained solely on human-annotated examples could eas-

ily over-fit specific question styles or image contents that are being asked, leaving the model largely ignorant about the sheer diversity of questions. Existing methods address this issue primarily by introducing an auxiliary task such as visual grounding, cycle consistency, or debiasing. In this paper, we take a drastically different approach. We found that many of the "unknowns" to the learned VQA model are indeed "known" in the dataset implicitly. For instance, questions asking about the same object in different images are likely paraphrases; the number of detected or annotated objects in an image already provides the answer to the "how many" question, even if the question has not been annotated for that image. Building upon these insights, we present a simple data augmentation pipeline SimpleAug to turn this "known" knowledge into training examples for VQA. We show that these augmented examples can notably improve the learned VQA models' performance, not only on the VQA-CP dataset with language prior shifts but also on the VQA v2 dataset without such shifts. Our method further opens up the door to leverage weakly-labeled or unlabeled images in a principled way to enhance VQA models. Our code and data are publicly available at https://github.com/heendung/simpleAUG.

Improving Pre-trained Vision-and-Language Embeddings for Phrase Grounding

Zi-Yi Dou and Nanyun Peng

Phrase grounding aims to map textual phrases to their associated image regions, which can be a prerequisite for multimodal reasoning and can benefit tasks requiring identifying objects based on language. With pre-trained vision-and-language models achieving impressive performance across tasks, it remains unclear if we can directly utilize their learned embeddings for phrase grounding without fine-tuning. To this end, we propose a method to extract matched phrase-region pairs from pre-trained vision-and-language embeddings and propose four fine-tuning objectives to improve the model phrase grounding ability using image-caption data without any supervised grounding signals. Experiments on two representative datasets demonstrate the effectiveness of our objectives, outperforming baseline models in both weakly-supervised and supervised phrase grounding settings. In addition, we evaluate the aligned embeddings on several other downstream tasks and show that we can achieve better phrase grounding without sacrificing representation generality.

Sequential Randomized Smoothing for Adversarially Robust Speech Recognition

Raphael Olivier and Bhiksha Raj

While Automatic Speech Recognition has been shown to be vulnerable to adversarial attacks, defenses against these attacks are still lagging. Existing, naive defenses can be partially broken with an adaptive attack. In classification tasks, the Randomized Smoothing paradigm has been shown to be effective at defending models. However, it is difficult to apply this paradigm to ASR tasks, due to their complexity and the sequential nature of their outputs. Our paper overcomes some of these challenges by leveraging speech-specific tools like enhancement and ROVER voting to design an ASR model that is robust to perturbations. We apply adaptive versions of state-of-the-art attacks, such as the Imperceptible ASR attack, to our model, and show that our strongest defense is robust to all attacks that use inaudible noise, and can only be broken with very high distortion.

Hitting your MARQ: Multimodal ARgument Quality Assessment in Long Debate Video

Md Kamrul Hasan et al.

The combination of gestures, intonations, and textual content plays a key role in argument delivery. However, the current literature mostly considers textual content while assessing the quality of an argument, and it is limited to datasets containing short sequences (18-48 words). In this paper, we study argument quality assessment in a multimodal context, and experiment on DBATES, a publicly available dataset of long debate videos. First, we propose a set of interpretable debate centric features such as clarity, content variation, body movement cues, and pauses, inspired by theories of argumentation quality. Second, we design the Multimodal ARgument Quality assessor (MARQ) – a hierarchical neural network model that summarizes the multimodal signals on long sequences and enriches the multimodal embedding with debate centric features. Our proposed MARQ model achieves an accuracy of 81.91% on the argument quality prediction task and outperforms established baseline models with an error rate reduction of 22.7%. Through ablation studies, we demonstrate the importance of multimodal cues in modeling argument quality.

Mind the Context: The Impact of Contextualization in Neural Module Networks for Grounding Visual Referring Expressions

Arjun Akula et al.

Neural module networks (NMN) are a popular approach for grounding visual referring expressions. Prior implementations of NMN use pre-defined and fixed textual inputs in their module instantiation. This necessitates a large number of modules as they lack the ability to share weights and exploit associations between similar textual contexts (e.g. "dark cube on the left" vs. "black cube on the left"). In this work, we address these limitations and evaluate the impact of contextual clues in improving the performance of NMN models. First, we address the problem of fixed textual inputs by parameterizing the module arguments. This substantially reduce the number of modules in NMN by up to 75% without any loss in performance. Next we propose a method to contextualize our parameterized model to enhance the module's capacity in exploiting the visiolinguistic associations. Our model outperforms the state-of-the-art NMN model on CLEVR-Ref+ dataset with +8.1% improvement in accuracy on the single-referent test set and +4.3% on the full test set. Additionally, we demonstrate that contextualization provides +11.2% and +1.7% improvements in accuracy over prior NMN models on CLOSURE and NLVR2. We further evaluate the impact of our contextualization by constructing a contrast set for CLEVR-Ref+, which we call CC-Ref+. We significantly outperform the baselines by as much as +10.4% absolute accuracy on CC-Ref+, illustrating the generalization skills of our approach.

Weakly-Supervised Visual-Retriever-Reader for Knowledge-based Question Answering
Man Luo et al.

Knowledge-based visual question answering (VQA) requires answering questions with external knowledge in addition to the content of images. One dataset that is mostly used in evaluating knowledge-based VQA is OK-VQA, but it lacks a gold standard knowledge corpus for retrieval. Existing work leverage different knowledge bases (e.g., ConceptNet and Wikipedia) to obtain external knowledge. Because of varying knowledge bases, it is hard to fairly compare models' performance. To address this issue, we collect a natural language knowledge base that can be used for any VQA system. Moreover, we propose a Visual Retriever-Reader pipeline to approach knowledge-based VQA. The visual retriever aims to retrieve relevant knowledge, and the visual reader seeks to predict answers based on given knowledge. We introduce various ways to retrieve knowledge using text and images and two reader styles: classification and extraction. Both the retriever and reader are trained with weak supervision. Our experimental results show that a good retriever can significantly improve the reader's performance on the OK-VQA challenge. The code and corpus are provided in https://github.com/luomancs/retriever.

NDH-Full: Learning and Evaluating Navigational Agents on Full-Length Dialogue
Hyounghun Kim, Jialu Li, and Mohit Bansal

Communication between human and mobile agents is getting increasingly important as such agents are widely deployed in our daily lives. Vision-and-Dialogue Navigation is one of the tasks that evaluate the agent's ability to interact with humans for assistance and navigate based on natural language responses. In this paper, we explore the Navigation from Dialogue History (NDH) task, which is based on the Cooperative Vision-and-Dialogue Navigation (CVDN) dataset, and present a state-of-the-art model which is built upon Vision-Language transformers. However, despite achieving competitive performance, we find that the agent in the NDH task is not evaluated appropriately by the primary metric – Goal Progress. By analyzing the performance mismatch between Goal Progress and other metrics (e.g., normalized Dynamic Time Warping) from our state-of-the-art model, we show that NDH's sub-path based task setup (i.e., navigating partial trajectory based on its correspondent subset of the full dialogue) does not provide the agent with enough supervision signal towards the goal region. Therefore, we propose a new task setup called NDH-Full which takes the full dialogue and the whole navigation path as one instance. We present a strong baseline model and show initial results on this new task. We further describe several approaches that we try, in order to improve the model performance (based on curriculum learning, pre-training, and data-augmentation), suggesting potential useful training methods on this new NDH-Full task.

Timeline Summarization based on Event Graph Compression via Time-Aware Optimal Transport
Manling Li et al.

Timeline Summarization identifies major events from a news collection and describes them following temporal order, with key dates tagged. Previous methods generally generate summaries separately for each date after they determine the key dates of events. These methods overlook the events' intra-structures (arguments) and inter-structures (event-event connections). Following a different route, we propose to represent the news articles as an event-graph, thus the summarization becomes compressing the whole graph to its salient subgraph. The key hypothesis is that the events connected through shared arguments and temporal order depict the skeleton of a timeline, containing events that are semantically related, temporally coherent and structurally salient in the global event graph. A time-aware optimal transport distance is then introduced for learning the compression model in an unsupervised manner. We show that our approach significantly improves on the state of the art on three real-world datasets, including two public standard benchmarks and our newly collected Timeline100 dataset.

StreamHover: Livestream Transcript Summarization and Annotation
Sangwoo Cho et al.

With the explosive growth of livestream broadcasting, there is an urgent need for new summarization technology that enables us to create a preview of streamed content and tap into this wealth of knowledge. However, the problem is nontrivial due to the informal nature of spoken language. Further, there has been a shortage of annotated datasets that are necessary for transcript summarization. In this paper, we present StreamHover, a framework for annotating and summarizing livestream transcripts. With a total of over 500 hours of videos annotated with both extractive and abstractive summaries, our benchmark dataset is significantly larger than currently existing annotated corpora. We explore a neural extractive summarization model that leverages vector-quantized variational autoencoder to learn latent vector representations of spoken utterances and identify salient utterances from the transcripts to form summaries. We show that our model generalizes better and improves performance over strong baselines. The results of this study provide an avenue for future research to improve summarization solutions for efficient browsing of livestreams.

Cross-Register Projection for Headline Part of Speech Tagging
Adrian Benton, Hanyang Li, and Igor Malioutov

Part of speech (POS) tagging is a familiar NLP task. State of the art taggers routinely achieve token-level accuracies of over 97% on news body text, evidence that the problem is well understood. However, the register of English news headlines, "headlinese", is very different from the register of long-form text, causing POS tagging models to underperform on headlines. In this work, we automatically annotate news headlines with POS tags by projecting predicted tags from corresponding sentences in news bodies. We train a multi-domain

POS tagger on both long-form and headline text and show that joint training on both registers improves over training on just one or naïvely concatenating training sets. We evaluate on a newly-annotated corpus of over 5,248 English news headlines from the Google sentence compression corpus, and show that our model yields a 23% relative error reduction per token and 19% per headline. In addition, we demonstrate that better headline POS tags can improve the performance of a syntax-based open information extraction system. We make POSH, the POS-tagged Headline corpus, available to encourage research in improved NLP models for news headlines.

Using Sociolinguistic Variables to Reveal Changing Attitudes Towards Sexuality and Gender
Sky CH-Wang and David Jurgens

Individuals signal aspects of their identity and beliefs through linguistic choices. Studying these choices in aggregate allows us to examine large-scale attitude shifts within a population. Here, we develop computational methods to study word choice within a sociolinguistic lexical variable—alternate words used to express the same concept—in order to test for change in the United States towards sexuality and gender. We examine two variables: i) referents to significant others, such as the word "partner" and ii) referents to an indefinite person, both of which could optionally be marked with gender. The linguistic choices in each variable allow us to study increased rates of acceptances of gay marriage and gender equality, respectively. In longitudinal analyses across Twitter and Reddit over 87M messages, we demonstrate that attitudes are changing but that these changes are driven by specific demographics within the United States. Further, in a quasi-causal analysis, we show that passages of Marriage Equality Acts in different states are drivers of linguistic change.

Idiosyncratic but not Arbitrary: Learning Idiolects in Online Registers Reveals Distinctive yet Consistent Individual Styles
Jian Zhu and David Jurgens

An individual's variation in writing style is often a function of both social and personal attributes. While structured social variation has been extensively studied, e.g., gender based variation, far less is known about how to characterize individual styles due to their idiosyncratic nature. We introduce a new approach to studying idiolects through a massive cross-author comparison to identify and encode stylistic features. The neural model achieves strong performance at authorship identification on short texts and through an analogy-based probing task, showing that the learned representations exhibit surprising regularities that encode qualitative and quantitative shifts of idiolectal styles. Through text perturbation, we quantify the relative contributions of different linguistic elements to idiolectal variation. Furthermore, we provide a description of idiolects through measuring inter- and intra-author variation, showing that variation in idiolects is often distinctive yet consistent.

Identifying Morality Frames in Political Tweets using Relational Learning
Shamik Roy, Maria Leonor Pacheco, and Dan Goldwasser

Extracting moral sentiment from text is a vital component in understanding public opinion, social movements, and policy decisions. The Moral Foundation Theory identifies five moral foundations, each associated with a positive and negative polarity. However, moral sentiment is often motivated by its targets, which can correspond to individuals or collective entities. In this paper, we introduce morality frames, a representation framework for organizing moral attitudes directed at different entities, and come up with a novel and high-quality annotated dataset of tweets written by US politicians. Then, we propose a relational learning model to predict moral attitudes towards entities and moral foundations jointly. We do qualitative and quantitative evaluations, showing that moral sentiment towards entities differs highly across political ideologies.

Measuring Sentence-Level and Aspect-Level Certainty in Science Communications
Jiaxin Pei and David Jurgens

Certainty and uncertainty are fundamental to science communication. Hedges have widely been used as proxies for uncertainty. However, certainty is a complex construct, with authors expressing not only the degree but the type and aspects of uncertainty in order to give the reader a certain impression of what is known. Here, we introduce a new study of certainty that models both the level and the aspects of certainty in scientific findings. Using a new dataset of 2167 annotated scientific findings, we demonstrate that hedges alone account for only a partial explanation of certainty. We show that both the overall certainty and individual aspects can be predicted with pre-trained language models, providing a more complete picture of the author's intended communication. Downstream analyses on 431K scientific findings from news and scientific abstracts demonstrate that modeling sentence-level and aspect-level certainty is meaningful for areas like science communication. Both the model and datasets used in this paper are released at https://blablablab.si.umich.edu/projects/certainty/.

(Mis)alignment Between Stance Expressed in Social Media Data and Public Opinion Surveys
Kenneth Joseph et al.

Stance detection, which aims to determine whether an individual is for or against a target concept, promises to uncover public opinion from large streams of social media data. Yet even human annotation of social media content does not always capture "stance" as measured by public opinion polls. We demonstrate this by directly comparing an individual's self-reported stance to the stance inferred from their social media data. Leveraging a longitudinal public opinion survey with respondent Twitter handles, we conducted this comparison for 1,129 individuals across four salient targets. We find that recall is high for both "Pro" and "Anti" stance classifications

but precision is variable in a number of cases. We identify three factors leading to the disconnect between text and author stance: temporal inconsistencies, differences in constructs, and measurement errors from both survey respondents and annotators. By presenting a framework for assessing the limitations of stance detection models, this work provides important insight into what stance detection truly measures.

"Was it "stated" or was it "claimed"?: How linguistic bias affects generative language models

Roma Patel and Ellie Pavlick

People use language in subtle and nuanced ways to convey their beliefs. For instance, saying *claimed* instead of *said* casts doubt on the truthfulness of the underlying proposition, thus representing the author's opinion on the matter. Several works have identified such linguistic classes of words that occur frequently in natural language text and are bias-inducing by virtue of their framing effects. In this paper, we test whether generative language models (including GPT-2 [**radford2019language**] are sensitive to these linguistic framing effects. In particular, we test whether prompts that contain linguistic markers of author bias (e.g., hedges, implicatives, subjective intensifiers, assertives) influence the distribution of the generated text. Although these framing effects are subtle and stylistic, we find evidence that they lead to measurable style and topic differences in the generated text, leading to language that is, on average, more polarised and more skewed towards controversial entities and events.

Latent Hatred: A Benchmark for Understanding Implicit Hate Speech

Mai ElSherief et al.

Hate speech has grown significantly on social media, causing serious consequences for victims of all demographics. Despite much attention being paid to characterize and detect discriminatory speech, most work has focused on explicit or overt hate speech, failing to address a more pervasive form based on coded or indirect language. To fill this gap, this work introduces a theoretically-justified taxonomy of implicit hate speech and a benchmark corpus with fine-grained labels for each message and its implication. We present systematic analyses of our dataset using contemporary baselines to detect and explain implicit hate speech, and we discuss key features that challenge existing models. This dataset will continue to serve as a useful benchmark for understanding this multifaceted issue.

GOLD: Improving Out-of-Scope Detection in Dialogues using Data Augmentation

Derek Chen and Zhou Yu

Practical dialogue systems require robust methods of detecting out-of-scope (OOS) utterances to avoid conversational breakdowns and related failure modes. Directly training a model with labeled OOS examples yields reasonable performance, but obtaining such data is a resource-intensive process. To tackle this limited-data problem, previous methods focus on better modeling the distribution of in-scope (INS) examples. We introduce GOLD as an orthogonal technique that augments existing data to train better OOS detectors operating in low-data regimes. GOLD generates pseudo-labeled candidates using samples from an auxiliary dataset and keeps only the most beneficial candidates for training through a novel filtering mechanism. In experiments across three target benchmarks, the top GOLD model outperforms all existing methods on all key metrics, achieving relative gains of 52.4%, 48.9% and 50.3% against median baseline performance. We also analyze the unique properties of OOS data to identify key factors for optimally applying our proposed method.

Towards Automatic Evaluation of Dialog Systems: A Model-Free Off-Policy Evaluation Approach

Haoming Jiang et al.

Reliable automatic evaluation of dialogue systems under an interactive environment has long been overdue. An ideal environment for evaluating dialog systems, also known as the Turing test, needs to involve human interaction, which is usually not affordable for large-scale experiments. Though researchers have attempted to use metrics for language generation tasks (e.g., perplexity, BLEU) or some model-based reinforcement learning methods (e.g., self-play evaluation) for automatic evaluation, these methods only show very weak correlation with the actual human evaluation in practice. To bridge such a gap, we propose a new framework named ENIGMA for estimating human evaluation scores based on recent advances of off-policy evaluation in reinforcement learning. ENIGMA only requires a handful of pre-collected experience data, and therefore does not involve human interaction with the target policy during the evaluation, making automatic evaluations feasible. More importantly, ENIGMA is model-free and agnostic to the behavior policies for collecting the experience data, which significantly alleviates the technical difficulties of modeling complex dialogue environments and human behaviors. Our experiments show that ENIGMA significantly outperforms existing methods in terms of correlation with human evaluation scores.

Automatically Exposing Problems with Neural Dialog Models

Dian Yu and Kenji Sagae

Neural dialog models are known to suffer from problems such as generating unsafe and inconsistent responses. Even though these problems are crucial and prevalent, they are mostly manually identified by model designers through interactions. Recently, some research instructs crowdworkers to goad the bots into triggering such problems. However, humans leverage superficial clues such as hate speech, while leaving systematic problems undercover. In this paper, we propose two methods including reinforcement learning to automatically trigger a dialog model into generating problematic responses. We show the effect of our methods in exposing safety

and contradiction issues with state-of-the-art dialog models.

MindCraft: Theory of Mind Modeling for Situated Dialogue in Collaborative Tasks
Cristian-Paul Bara, Sky CH-Wang, and Joyce Chai

An ideal integration of autonomous agents in a human world implies that they are able to collaborate on human terms. In particular, theory of mind plays an important role in maintaining common ground during human collaboration and communication. To enable theory of mind modeling in situated interactions, we introduce a fine-grained dataset of collaborative tasks performed by pairs of human subjects in the 3D virtual blocks world of Minecraft. It provides information that captures partners' beliefs of the world and of each other as an interaction unfolds, bringing abundant opportunities to study human collaborative behaviors in situated language communication. As a first step towards our goal of developing embodied AI agents able to infer belief states of collaborative partners in situ, we build and present results on computational models for several theory of mind tasks.

Contextual Rephrase Detection for Reducing Friction in Dialogue Systems
Zhuoyi Wang et al.

For voice assistants like Alexa, Google Assistant, and Siri, correctly interpreting users' intentions is of utmost importance. However, users sometimes experience friction with these assistants, caused by errors from different system components or user errors such as slips of the tongue. Users tend to rephrase their queries until they get a satisfactory response. Rephrase detection is used to identify the rephrases and has long been treated as a task with pairwise input, which does not fully utilize the contextual information (e.g. users' implicit feedback). To this end, we propose a contextual rephrase detection model ContReph to automatically identify rephrases from multi-turn dialogues. We showcase how to leverage the dialogue context and user-agent interaction signals, including the user's implicit feedback and the time gap between different turns, which can help significantly outperform the pairwise rephrase detection models.

Few-Shot Intent Detection via Contrastive Pre-Training and Fine-Tuning
Jianguo Zhang et al.

In this work, we focus on a more challenging few-shot intent detection scenario where many intents are fine-grained and semantically similar. We present a simple yet effective few-shot intent detection schema via contrastive pre-training and fine-tuning. Specifically, we first conduct self-supervised contrastive pre-training on collected intent datasets, which implicitly learns to discriminate semantically similar utterances without using any labels. We then perform few-shot intent detection together with supervised contrastive learning, which explicitly pulls utterances from the same intent closer and pushes utterances across different intents farther. Experimental results show that our proposed method achieves state-of-the-art performance on three challenging intent detection datasets under 5-shot and 10-shot settings.

DIALKI: Knowledge Identification in Conversational Systems through Dialogue-Document Contextualization
Zeqiu Wu et al.

Identifying relevant knowledge to be used in conversational systems that are grounded in long documents is critical to effective response generation. We introduce a knowledge identification model that leverages the document structure to provide dialogue-contextualized passage encodings and better locate knowledge relevant to the conversation. An auxiliary loss captures the history of dialogue-document connections. We demonstrate the effectiveness of our model on two document-grounded conversational datasets and provide analyses showing generalization to unseen documents and long dialogue contexts.

Investigating Robustness of Dialog Models to Popular Figurative Language Constructs
Harsh Jhamtani et al.

Humans often employ figurative language use in communication, including during interactions with dialog systems. Thus, it is important for real-world dialog systems to be able to handle popular figurative language constructs like metaphor and simile. In this work, we analyze the performance of existing dialog models in situations where the input dialog context exhibits use of figurative language. We observe large gaps in handling of figurative language when evaluating the models on two open domain dialog datasets. When faced with dialog contexts consisting of figurative language, some models show very large drops in performance compared to contexts without figurative language. We encourage future research in dialog modeling to separately analyze and report results on figurative language in order to better test model capabilities relevant to real-world use. Finally, we propose lightweight solutions to help existing models become more robust to figurative language by simply using an external resource to translate figurative language to literal (non-figurative) forms while preserving the meaning to the best extent possible.

"It doesn't look good for a date": Transforming Critiques into Preferences for Conversational Recommendation Systems
Victor Bursztyn et al.

Conversations aimed at determining good recommendations are iterative in nature. People often express their preferences in terms of a critique of the current recommendation (e.g., "It doesn't look good for a date"), requiring some degree of common sense for a preference to be inferred. In this work, we present a method

for transforming a user critique into a positive preference (e.g., "I prefer more romantic") in order to retrieve reviews pertaining to potentially better recommendations (e.g., "Perfect for a romantic dinner"). We leverage a large neural language model (LM) in a few-shot setting to perform critique-to-preference transformation, and we test two methods for retrieving recommendations: one that matches embeddings, and another that fine-tunes an LM for the task. We instantiate this approach in the restaurant domain and evaluate it using a new dataset of restaurant critiques. In an ablation study, we show that utilizing critique-to-preference transformation improves recommendations, and that there are at least three general cases that explain this improved performance.

Iconary: A Pictionary-Based Game for Testing Multimodal Communication with Drawings and Text
Christopher Clark et al.

Communicating with humans is challenging for AIs because it requires a shared understanding of the world, complex semantics (e.g., metaphors or analogies), and at times multi-modal gestures (e.g., pointing with a finger, or an arrow in a diagram). We investigate these challenges in the context of Iconary, a collaborative game of drawing and guessing based on Pictionary, that poses a novel challenge for the research community. In Iconary, a Guesser tries to identify a phrase that a Drawer is drawing by composing icons, and the Drawer iteratively revises the drawing to help the Guesser in response. This back-and-forth often uses canonical scenes, visual metaphor, or icon compositions to express challenging words, making it an ideal test for mixing language and visual/symbolic communication in AI. We propose models to play Iconary and train them on over 55,000 games between human players. Our models are skillful players and are able to employ world knowledge in language models to play with words unseen during training.

Effective Sequence-to-Sequence Dialogue State Tracking
Jeffrey Zhao et al.

Sequence-to-sequence models have been applied to a wide variety of NLP tasks, but how to properly use them for dialogue state tracking has not been systematically investigated. In this paper, we study this problem from the perspectives of pre-training objectives as well as the formats of context representations. We demonstrate that the choice of pre-training objective makes a significant difference to the state tracking quality. In particular, we find that masked span prediction is more effective than auto-regressive language modeling. We also explore using Pegasus, a span prediction-based pre-training objective for text summarization, for the state tracking model. We found that pre-training for the seemingly distant summarization task works surprisingly well for dialogue state tracking. In addition, we found that while recurrent state context representation works also reasonably well, the model may have a hard time recovering from earlier mistakes. We conducted experiments on the MultiWOZ 2.1-2.4, WOZ 2.0, and DSTC2 datasets with consistent observations.

Understanding Politics via Contextualized Discourse Processing
Rajkumar Pujari and Dan Goldwasser

Politicians often have underlying agendas when reacting to events. Arguments in contexts of various events reflect a fairly consistent set of agendas for a given entity. In spite of recent advances in Pretrained Language Models, those text representations are not designed to capture such nuanced patterns. In this paper, we propose a Compositional Reader model consisting of encoder and composer modules, that captures and leverages such information to generate more effective representations for entities, issues, and events. These representations are contextualized by tweets, press releases, issues, news articles, and participating entities. Our model processes several documents at once and generates composed representations for multiple entities over several issues or events. Via qualitative and quantitative empirical analysis, we show that these representations are meaningful and effective.

Inducing Stereotypical Character Roles from Plot Structure
Labiba Jahan, Rahul Mittal, and Mark Finlayson

Stereotypical character roles-also known as archetypes or dramatis personae-play an important function in narratives: they facilitate efficient communication with bundles of default characteristics and associations and ease understanding of those characters' roles in the overall narrative. We present a fully unsupervised k-means clustering approach for learning stereotypical roles given only structural plot information. We demonstrate the technique on Vladimir Propp's structural theory of Russian folktales (captured in the extended ProppLearner corpus, with 46 tales), showing that our approach can induce six out of seven of Propp's dramatis personae with F1 measures of up to 0.70 (0.58 average), with an additional category for minor characters. We have explored various feature sets and variations of a cluster evaluation method. The best-performing feature set comprises plot functions, unigrams, tf-idf weights, and embeddings over coreference chain heads. Roles that are mentioned more often (Hero, Villain), or have clearly distinct plot patterns (Princess) are more strongly differentiated than less frequent or distinct roles (Dispatcher, Helper, Donor). Detailed error analysis suggests that the quality of the coreference chain and plot functions annotations are critical for this task. We provide all our data and code for reproducibility.

Conundrums in Event Coreference Resolution: Making Sense of the State of the Art
Jing Lu and Vincent Ng

Despite recent promising results on the application of span-based models for event reference interpretation, there is a lack of understanding of what has been improved. We present an empirical analysis of a state-of-the-art span-based event reference systems with the goal of providing the general NLP audience with a better

understanding of the state of the art and reference researchers with directions for future research.

Focus on what matters: Applying Discourse Coherence Theory to Cross Document Coreference
William Held, Dan Iter, and Dan Jurafsky

Performing event and entity coreference resolution across documents vastly increases the number of candidate mentions, making it intractable to do the full n^2 pairwise comparisons. Existing approaches simplify by considering coreference only within document clusters, but this fails to handle inter-cluster coreference, common in many applications. As a result cross-document coreference algorithms are rarely applied to downstream tasks. We draw on an insight from discourse coherence theory: potential coreferences are constrained by the reader's discourse focus. We model the entities/events in a reader's focus as a neighborhood within a learned latent embedding space which minimizes the distance between mentions and the centroids of their gold coreference clusters. We then use these neighborhoods to sample only hard negatives to train a fine-grained classifier on mention pairs and their local discourse features. Our approach achieves state-of-the-art results for both events and entities on the ECB+, Gun Violence, Football Coreference, and Cross-Domain Cross-Document Coreference corpora. Furthermore, training on multiple corpora improves average performance across all datasets by 17.2 F1 points, leading to a robust coreference resolution model that is now feasible to apply to downstream tasks.

Salience-Aware Event Chain Modeling for Narrative Understanding
Xiyang Zhang, Muhao Chen, and Jonathan May

Storytelling, whether via fables, news reports, documentaries, or memoirs, can be thought of as the communication of interesting and related events that, taken together, form a concrete process. It is desirable to extract the event chains that represent such processes. However, this extraction remains a challenging problem. We posit that this is due to the nature of the texts from which chains are discovered. Natural language text interleaves a narrative of concrete, salient events with background information, contextualization, opinion, and other elements that are important for a variety of necessary discourse and pragmatics acts but are not part of the principal chain of events being communicated. We introduce methods for extracting this principal chain from natural language text, by filtering away non-salient events and supportive sentences. We demonstrate the effectiveness of our methods at isolating critical event chains by comparing their effect on downstream tasks. We show that by pre-training large language models on our extracted chains, we obtain improvements in two tasks that benefit from a clear understanding of event chains: narrative prediction and event-based temporal question answering. The demonstrated improvements and ablative studies confirm that our extraction method isolates critical event chains.

Multitask Semi-Supervised Learning for Class-Imbalanced Discourse Classification
Alexander Spangher et al.

As labeling schemas evolve over time, small differences can render datasets following older schemas unusable. This prevents researchers from building on top of previous annotation work and results in the existence, in discourse learning in particular, of many small class-imbalanced datasets. In this work, we show that a multitask learning approach can combine discourse datasets from similar and diverse domains to improve discourse classification. We show an improvement of 4.9% Micro F1-score over current state-of-the-art benchmarks on the *NewsDiscourse* dataset, one of the largest discourse datasets recently published, due in part to label correlations across tasks, which improve performance for underrepresented classes. We also offer an extensive review of additional techniques proposed to address resource-poor problems in NLP, and show that none of these approaches can improve classification accuracy in our setting.

Finetuning Pretrained Transformers into RNNs
Jungo Kasai et al.

Transformers have outperformed recurrent neural networks (RNNs) in natural language generation. But this comes with a signifi- cant computational cost, as the attention mechanism's complexity scales quadratically with sequence length. Efficient transformer variants have received increasing interest in recent works. Among them, a linear-complexity recurrent variant has proven well suited for autoregressive generation. It approximates the softmax attention with randomized or heuristic feature maps, but can be difficult to train and may yield suboptimal accuracy. This work aims to convert a pretrained transformer into its efficient recurrent counterpart, improving efficiency while maintaining accuracy. Specifically, we propose a swap-then-finetune procedure: in an off-the-shelf pretrained transformer, we replace the softmax attention with its linear-complexity recurrent alternative and then finetune. With a learned feature map, our approach provides an improved trade-off between efficiency and accuracy over the standard transformer and other recurrent variants. We also show that the finetuning process has lower training cost relative to training these recurrent variants from scratch. As many models for natural language tasks are increasingly dependent on large-scale pretrained transformers, this work presents a viable approach to improving inference efficiency without repeating the expensive pretraining process.

Learning with Different Amounts of Annotation: From Zero to Many Labels
Shujian Zhang, Chengyue Gong, and Eunsol Choi

Training NLP systems typically assumes access to annotated data that has a single human label per example. Given imperfect labeling from annotators and inherent ambiguity of language, we hypothesize that single

label is not sufficient to learn the spectrum of language interpretation. We explore new annotation distribution schemes, assigning multiple labels per example for a small subset of training examples. Introducing such multi label examples at the cost of annotating fewer examples brings clear gains on natural language inference task and entity typing task, even when we simply first train with a single label data and then fine tune with multi label examples. Extending a MixUp data augmentation framework, we propose a learning algorithm that can learn from training examples with different amount of annotation (with zero, one, or multiple labels). This algorithm efficiently combines signals from uneven training data and brings additional gains in low annotation budget and cross domain settings. Together, our method achieves consistent gains in two tasks, suggesting distributing labels unevenly among training examples can be beneficial for many NLP tasks.

Beyond Preserved Accuracy: Evaluating Loyalty and Robustness of BERT Compression
Canwen Xu et al.

Recent studies on compression of pretrained language models (e.g., BERT) usually use preserved accuracy as the metric for evaluation. In this paper, we propose two new metrics, label loyalty and probability loyalty that measure how closely a compressed model (i.e., student) mimics the original model (i.e., teacher). We also explore the effect of compression with regard to robustness under adversarial attacks. We benchmark quantization, pruning, knowledge distillation and progressive module replacing with loyalty and robustness. By combining multiple compression techniques, we provide a practical strategy to achieve better accuracy, loyalty and robustness.

A Secure and Efficient Federated Learning Framework for NLP
Chenghong Wang et al.

In this work, we consider the problem of designing secure and efficient federated learning (FL) frameworks for NLP. Existing solutions under this literature either consider a trusted aggregator or require heavy-weight cryptographic primitives, which makes the performance significantly degraded. Moreover, many existing secure FL designs work only under the restrictive assumption that none of the clients can be dropped out from the training protocol. To tackle these problems, we propose SEFL, a secure and efficient federated learning framework that (1)~eliminates the need for the trusted entities; (2)~achieves similar and even better model accuracy compared with existing FL designs; (3)~is resilient to client dropouts.

Reconstruction Attack on Instance Encoding for Language Understanding
Shangyu Xie and Yuan Hong

A private learning scheme TextHide was recently proposed to protect the private text data during the training phase via so-called instance encoding. We propose a novel reconstruction attack to break TextHide by recovering the private training data, and thus unveil the privacy risks of instance encoding. We have experimentally validated the effectiveness of the reconstruction attack with two commonly-used datasets for sentence classification. Our attack would advance the development of privacy preserving machine learning in the context of natural language processing.

Style Pooling: Automatic Text Style Obfuscation for Improved Classification Fairness
Fatemehsadat Mireshghallah and Taylor Berg-Kirkpatrick

Text style can reveal sensitive attributes of the author (e.g. age and race) to the reader, which can, in turn, lead to privacy violations and bias in both human and algorithmic decisions based on text. For example, the style of writing in job applications might reveal protected attributes of the candidate which could lead to bias in hiring decisions, regardless of whether hiring decisions are made algorithmically or by humans. We propose a VAE-based framework that obfuscates stylistic features of human-generated text through style transfer, by automatically re-writing the text itself. Critically, our framework operationalizes the notion of obfuscated style in a flexible way that enables two distinct notions of obfuscated style: (1) a minimal notion that effectively intersects the various styles seen in training, and (2) a maximal notion that seeks to obfuscate by adding stylistic features of all sensitive attributes to text, in effect, computing a union of styles. Our style-obfuscation framework can be used for multiple purposes, however, we demonstrate its effectiveness in improving the fairness of downstream classifiers. We also conduct a comprehensive study on style-pooling's effect on fluency, semantic consistency, and attribute removal from text, in two and three domain style transfer.

Low Frequency Names Exhibit Bias and Overfitting in Contextualizing Language Models
Robert Wolfe and Aylin Caliskan

We use a dataset of U.S. first names with labels based on predominant gender and racial group to examine the effect of training corpus frequency on tokenization, contextualization, similarity to initial representation, and bias in BERT, GPT-2, T5, and XLNet. We show that predominantly female and non-white names are less frequent in the training corpora of these four language models. We find that infrequent names are more self-similar across contexts, with Spearman's rho between frequency and self-similarity as low as -.763. Infrequent names are also less similar to initial representation, with Spearman's rho between frequency and linear centered kernel alignment (CKA) similarity to initial representation as high as .702. Moreover, we find Spearman's rho between racial bias and name frequency in BERT of .492, indicating that lower-frequency minority group names are more associated with unpleasantness. Representations of infrequent names undergo more processing, but are more self-similar, indicating that models rely on less context-informed representations of uncommon and minority names which are overfit to a lower number of observed contexts.

Adversarial Scrubbing of Demographic Information for Text Classification
Somnath Basu Roy Chowdhury et al.

Contextual representations learned by language models can often encode undesirable attributes, like demographic associations of the users, while being trained for an unrelated target task. We aim to scrub such undesirable attributes and learn fair representations while maintaining performance on the target task. In this paper, we present an adversarial learning framework "Adversarial Scrubber" (AdS), to debias contextual representations. We perform theoretical analysis to show that our framework converges without leaking demographic information under certain conditions. We extend previous evaluation techniques by evaluating debiasing performance using Minimum Description Length (MDL) probing. Experimental evaluations on 8 datasets show that AdS generates representations with minimal information about demographic attributes while being maximally informative about the target task.

CRYPTOGRU: Low Latency Privacy-Preserving Text Analysis With GRU
Bo Feng et al.

Homomorphic encryption (HE) and garbled circuit (GC) provide the protection for users' privacy. However, simply mixing the HE and GC in RNN models suffer from long inference latency due to slow activation functions. In this paper, we present a novel hybrid structure of HE and GC gated recurrent unit (GRU) network, , for low-latency secure inferences. replaces computationally expensive GC-based $tanh$ with fast GC-based $ReLU$, and then quantizes $sigmoid$ and $ReLU$ to smaller bit-length to accelerate activations in a GRU. We evaluate with multiple GRU models trained on 4 public datasets. Experimental results show achieves top-notch accuracy and improves the secure inference latency by up to $138\times$ over one of the state-of-the-art secure networks on the Penn Treebank dataset.

Modeling Disclosive Transparency in NLP Application Descriptions
Michael Saxon et al.

Broader disclosive transparency—truth and clarity in communication regarding the function of AI systems—is widely considered desirable. Unfortunately, it is a nebulous concept, difficult to both define and quantify. This is problematic, as previous work has demonstrated possible trade-offs and negative consequences to disclosive transparency, such as a confusion effect, where "too much information" clouds a reader's understanding of what a system description means. Disclosive transparency's subjective nature has rendered deep study into these problems and their remedies difficult. To improve this state of affairs, We introduce neural language model-based probabilistic metrics to directly model disclosive transparency, and demonstrate that they correlate with user and expert opinions of system transparency, making them a valid objective proxy. Finally, we demonstrate the use of these metrics in a pilot study quantifying the relationships between transparency, confusion, and user perceptions in a corpus of real NLP system descriptions.

OSCaR: Orthogonal Subspace Correction and Rectification of Biases in Word Embeddings
Sunipa Dev et al.

Language representations are known to carry stereotypical biases and, as a result, lead to biased predictions in downstream tasks. While existing methods are effective at mitigating biases by linear projection, such methods are too aggressive: they not only remove bias, but also erase valuable information from word embeddings. We develop new measures for evaluating specific information retention that demonstrate the tradeoff between bias removal and information retention. To address this challenge, we propose OSCaR (Orthogonal Subspace Correction and Rectification), a bias-mitigating method that focuses on disentangling biased associations between concepts instead of removing concepts wholesale. Our experiments on gender biases show that OSCaR is a well-balanced approach that ensures that semantic information is retained in the embeddings and bias is also effectively mitigated.

Injecting Entity Types into Entity-Guided Text Generation
Xiangyu Dong et al.

Recent successes in deep generative modeling have led to significant advances in natural language generation (NLG). Incorporating entities into neural generation models has demonstrated great improvements by assisting to infer the summary topic and to generate coherent content. To enhance the role of entity in NLG, in this paper, we aim to model the entity type in the decoding phase to generate contextual words accurately. We develop a novel NLG model to produce a target sequence based on a given list of entities. Our model has a multi-step decoder that injects the entity types into the process of entity mention generation. Experiments on two public news datasets demonstrate type injection performs better than existing type embedding concatenation baselines.

Data-to-text Generation by Splicing Together Nearest Neighbors
Sam Wiseman, Arturs Backurs, and Karl Stratos

We propose to tackle data-to-text generation tasks by directly splicing together retrieved segments of text from "neighbor" source-target pairs. Unlike recent work that conditions on retrieved neighbors but generates text token-by-token, left-to-right, we learn a policy that directly manipulates segments of neighbor text, by inserting or replacing them in partially constructed generations. Standard techniques for training such a policy require an oracle derivation for each generation, and we prove that finding the shortest such derivation can be reduced to parsing under a particular weighted context-free grammar. We find that policies learned in this way perform on par with strong baselines in terms of automatic and human evaluation, but allow for more interpretable and controllable generation.

Truth-Conditional Captions for Time Series Data
Harsh Jhamtani and Taylor Berg-Kirkpatrick

In this paper, we explore the task of automatically generating natural language descriptions of salient patterns in a time series, such as stock prices of a company over a week. A model for this task should be able to extract high-level patterns such as presence of a peak or a dip. While typical contemporary neural models with attention mechanisms can generate fluent output descriptions for this task, they often generate factually incorrect descriptions. We propose a computational model with a truth-conditional architecture which first runs small learned programs on the input time series, then identifies the programs/patterns which hold true for the given input, and finally conditions on *only* the chosen valid program (rather than the input time series) to generate the output text description. A program in our model is constructed from modules, which are small neural networks that are designed to capture numerical patterns and temporal information. The modules are shared across multiple programs, enabling compositionality as well as efficient learning of module parameters. The modules, as well as the composition of the modules, are unobserved in data, and we learn them in an end-to-end fashion with the only training signal coming from the accompanying natural language text descriptions. We find that the proposed model is able to generate high-precision captions even though we consider a small and simple space of module types.

Learning Compact Metrics for MT
Amy Pu et al.

Recent developments in machine translation and multilingual text generation have led researchers to adopt trained metrics such as COMET or BLEURT, which treat evaluation as a regression problem and use representations from multilingual pre-trained models such as XLM-RoBERTa or mBERT. Yet studies on related tasks suggest that these models are most efficient when they are large, which is costly and impractical for evaluation. We investigate the trade-off between multilinguality and model capacity with RemBERT, a state-of-the-art multilingual language model, using data from the WMT Metrics Shared Task. We present a series of experiments which show that model size is indeed a bottleneck for cross-lingual transfer, then demonstrate how distillation can help addressing this bottleneck, by leveraging synthetic data generation and transferring knowledge from one teacher to multiple students trained on related languages. Our method yields up to 10.5% improvement over vanilla fine-tuning and reaches 92.6% of RemBERT's performance using only a third of its parameters.

Back to the Basics: A Quantitative Analysis of Statistical and Graph-Based Term Weighting Schemes for Keyword Extraction
Asahi Ushio, Federico Liberatore, and Jose Camacho-Collados

Term weighting schemes are widely used in Natural Language Processing and Information Retrieval. In particular, term weighting is the basis for keyword extraction. However, there are relatively few evaluation studies that shed light about the strengths and shortcomings of each weighting scheme. In fact, in most cases researchers and practitioners resort to the well-known tf-idf as default, despite the existence of other suitable alternatives, including graph-based models. In this paper, we perform an exhaustive and large-scale empirical comparison of both statistical and graph-based term weighting methods in the context of keyword extraction. Our analysis reveals some interesting findings such as the advantages of the less-known lexical specificity with respect to tf-idf, or the qualitative differences between statistical and graph-based methods. Finally, based on our findings we discuss and devise some suggestions for practitioners. Source code to reproduce our experimental results, including a keyword extraction library, are available in the following repository: https://github.com/asahi417/kex

AttentionRank: Unsupervised Keyphrase Extraction using Self and Cross Attentions
Haoran Ding and Xiao Luo

Keyword or keyphrase extraction is to identify words or phrases presenting the main topics of a document. This paper proposes the AttentionRank, a hybrid attention model, to identify keyphrases from a document in an unsupervised manner. AttentionRank calculates self-attention and cross-attention using a pre-trained language model. The self-attention is designed to determine the importance of a candidate within the context of a sentence. The cross-attention is calculated to identify the semantic relevance between a candidate and sentences within a document. We evaluate the AttentionRank on three publicly available datasets against seven baselines. The results show that the AttentionRank is an effective and robust unsupervised keyphrase extraction model on both long and short documents. Source code is available on Github.

Open Knowledge Graphs Canonicalization using Variational Autoencoders
Sarthak Dash et al.

Noun phrases and Relation phrases in open knowledge graphs are not canonicalized, leading to an explosion of redundant and ambiguous subject-relation-object triples. Existing approaches to solve this problem take a two-step approach. First, they generate embedding representations for both noun and relation phrases, then a clustering algorithm is used to group them using the embeddings as features. In this work, we propose Canonicalizing Using Variational AutoEncoders and Side Information (CUVA), a joint model to learn both embeddings and cluster assignments in an end-to-end approach, which leads to a better vector representation for the noun and relation phrases. Our evaluation over multiple benchmarks shows that CUVA outperforms the existing state-of-the-art approaches. Moreover, we introduce CanonicNell, a novel dataset to evaluate entity canonicalization systems.

XLEnt: Mining a Large Cross-lingual Entity Dataset with Lexical-Semantic-Phonetic Word Alignment
Ahmed El-Kishky et al.

Cross-lingual named-entity lexica are an important resource to multilingual NLP tasks such as machine translation and cross-lingual wikification. While knowledge bases contain a large number of entities in high-resource languages such as English and French, corresponding entities for lower-resource languages are often missing. To address this, we propose Lexical-Semantic-Phonetic Align (LSP-Align), a technique to automatically mine cross-lingual entity lexica from mined web data. We demonstrate LSP-Align outperforms baselines at extracting cross-lingual entity pairs and mine 164 million entity pairs from 120 different languages aligned with English. We release these cross-lingual entity pairs along with the massively multilingual tagged named entity corpus as a resource to the NLP community.

Unsupervised Relation Extraction: A Variational Autoencoder Approach
Chenhan Yuan and Hoda Eldardiry

Unsupervised relation extraction works by clustering entity pairs that have the same relations in the text. Some existing variational autoencoder (VAE)-based approaches train the relation extraction model as an encoder that generates relation classifications. A decoder is trained along with the encoder to reconstruct the encoder input based on the encoder-generated relation classifications. These classifications are a latent variable so they are required to follow a pre-defined prior distribution which results in unstable training. We propose a VAE-based unsupervised relation extraction technique that overcomes this limitation by using the classifications as an intermediate variable instead of a latent variable. Specifically, classifications are conditioned on sentence input, while the latent variable is conditioned on both the classifications and the sentence input. This allows our model to connect the decoder with the encoder without putting restrictions on the classification distribution; which improves training stability. Our approach is evaluated on the NYT dataset and outperforms state-of-the-art methods.

Robust Retrieval Augmented Generation for Zero-shot Slot Filling
Michael Glass et al.

Automatically inducing high quality knowledge graphs from a given collection of documents still remains a challenging problem in AI. One way to make headway for this problem is through advancements in a related task known as slot filling. In this task, given an entity query in form of [Entity, Slot, ?], a system is asked to 'fill' the slot by generating or extracting the missing value exploiting evidence extracted from relevant passage(s) in the given document collection. The recent works in the field try to solve this task in an end-to-end fashion using retrieval-based language models. In this paper, we present a novel approach to zero-shot slot filling that extends dense passage retrieval with hard negatives and robust training procedures for retrieval augmented generation models. Our model reports large improvements on both T-REx and zsRE slot filling datasets, improving both passage retrieval and slot value generation, and ranking at the top-1 position in the KILT leaderboard. Moreover, we demonstrate the robustness of our system showing its domain adaptation capability on a new variant of the TACRED dataset for slot filling, through a combination of zero/few-shot learning. We release the source code and pre-trained models.

Everything Is All It Takes: A Multipronged Strategy for Zero-Shot Cross-Lingual Information Extraction
Mahsa Yarmohammadi et al.

Zero-shot cross-lingual information extraction (IE) describes the construction of an IE model for some target language, given existing annotations exclusively in some other language, typically English. While the advance of pretrained multilingual encoders suggests an easy optimism of "train on English, run on any language", we find through a thorough exploration and extension of techniques that a combination of approaches, both new and old, leads to better performance than any one cross-lingual strategy in particular. We explore techniques including data projection and self-training, and how different pretrained encoders impact them. We use English-to-Arabic IE as our initial example, demonstrating strong performance in this setting for event extraction, named entity recognition, part-of-speech tagging, and dependency parsing. We then apply data projection and self-training to three tasks across eight target languages. Because no single set of techniques performs the best across all tasks, we encourage practitioners to explore various configurations of the techniques described in this work when seeking to improve on zero-shot training.

HittER: Hierarchical Transformers for Knowledge Graph Embeddings
Sanxing Chen et al.

This paper examines the challenging problem of learning representations of entities and relations in a complex multi-relational knowledge graph. We propose HittER, a Hierarchical Transformer model to jointly learn Entity-relation composition and Relational contextualization based on a source entity's neighborhood. Our proposed model consists of two different Transformer blocks: the bottom block extracts features of each entity-relation pair in the local neighborhood of the source entity and the top block aggregates the relational information from outputs of the bottom block. We further design a masked entity prediction task to balance information from the relational context and the source entity itself. Experimental results show that HittER achieves new state-of-the-art results on multiple link prediction datasets. We additionally propose a simple approach to integrate HittER into BERT and demonstrate its effectiveness on two Freebase factoid question answering datasets.

Extracting Fine-Grained Knowledge Graphs of Scientific Claims: Dataset and Transformer-Based Results
Ian Magnusson and Scott Friedman

Recent transformer-based approaches demonstrate promising results on relational scientific information extraction. Existing datasets focus on high-level description of how research is carried out. Instead we focus on the subtleties of how experimental associations are presented by building SciClaim, a dataset of scientific claims drawn from Social and Behavior Science (SBS), PubMed, and CORD-19 papers. Our novel graph annotation schema incorporates not only coarse-grained entity spans as nodes and relations as edges between them, but also fine-grained attributes that modify entities and their relations, for a total of 12,738 labels in the corpus. By including more label types and more than twice the label density of previous datasets, SciClaim captures causal, comparative, predictive, statistical, and proportional associations over experimental variables along with their qualifications, subtypes, and evidence. We extend work in transformer-based joint entity and relation extraction to effectively infer our schema, showing the promise of fine-grained knowledge graphs in scientific claims and beyond.

Few-Shot Named Entity Recognition: An Empirical Baseline Study
Jiaxin Huang et al.

This paper presents an empirical study to efficiently build named entity recognition (NER) systems when a small amount of in-domain labeled data is available. Based upon recent Transformer-based self-supervised pre-trained language models (PLMs), we investigate three orthogonal schemes to improve model generalization ability in few-shot settings: (1) meta-learning to construct prototypes for different entity types, (2) task-specific supervised pre-training on noisy web data to extract entity-related representations and (3) self-training to leverage unlabeled in-domain data. On 10 public NER datasets, we perform extensive empirical comparisons over the proposed schemes and their combinations with various proportions of labeled data, our experiments show that (i)in the few-shot learning setting, the proposed NER schemes significantly improve or outperform the commonly used baseline, a PLM-based linear classifier fine-tuned using domain labels. (ii) We create new state-of-the-art results on both few-shot and training-free settings compared with existing methods.

Sequential Cross-Document Coreference Resolution
Emily Allaway, Shuai Wang, and Miguel Ballesteros

Relating entities and events in text is a key component of natural language understanding. Cross-document coreference resolution, in particular, is important for the growing interest in multi-document analysis tasks. In this work we propose a new model that extends the efficient sequential prediction paradigm for coreference resolution to cross-document settings and achieves competitive results for both entity and event coreference while providing strong evidence of the efficacy of both sequential models and higher-order inference in cross-document settings. Our model incrementally composes mentions into cluster representations and predicts links between a mention and the already constructed clusters, approximating a higher-order model. In addition, we conduct extensive ablation studies that provide new insights into the importance of various inputs and representation types in coreference.

Utilizing Relative Event Time to Enhance Event-Event Temporal Relation Extraction
Haoyang Wen and Heng Ji

Event time is one of the most important features for event-event temporal relation extraction. However, explicit event time information in text is sparse. For example, only about 20% of event mentions in TimeBank-Dense have event-time links. In this paper, we propose a joint model for event-event temporal relation classification and an auxiliary task, relative event time prediction, which predicts the event time as real numbers. We adopt the Stack-Propagation framework to incorporate predicted relative event time for temporal relation classification and keep the differentiability. Our experiments on MATRES dataset show that our model can significantly improve the RoBERTa-based baseline and achieve state-of-the-art performance.

Zero-Shot Information Extraction as a Unified Text-to-Triple Translation
Chenguang Wang et al.

We cast a suite of information extraction tasks into a text-to-triple translation framework. Instead of solving each task relying on task-specific datasets and models, we formalize the task as a translation between task-specific input text and output triples. By taking the task-specific input, we enable a task-agnostic translation by leveraging the latent knowledge that a pre-trained language model has about the task. We further demonstrate that a simple pre-training task of predicting which relational information corresponds to which input text is an effective way to produce task-specific outputs. This enables the zero-shot transfer of our framework to downstream tasks. We study the zero-shot performance of this framework on open information extraction (OIE2016, NYT, WEB, PENN), relation classification (FewRel and TACRED), and factual probe (Google-RE and T-REx). The model transfers non-trivially to most tasks and is often competitive with a fully supervised method without the need for any task-specific training. For instance, we significantly outperform the F1 score of the supervised open information extraction without needing to use its training set.

Partially Supervised Named Entity Recognition via the Expected Entity Ratio Loss
Thomas Effland and Michael Collins

We study learning named entity recognizers in the presence of missing entity annotations. We approach this setting as tagging with latent variables and propose a novel loss, the Expected Entity Ratio, to learn models

in the presence of systematically missing tags. We show that our approach is both theoretically sound and empirically useful. Experimentally, we find that it outperforms previous state-of-the-art baselines both in accuracy and speed across a variety of languages, annotation scenarios, and amounts of labeled data. In particular, we find that it outperforms the previous state of the art by $+12.7$ F1 score in a challenging setting with only $1,000$ biased annotations. We also show that, when combined with our approach, a novel "quick and sparse" annotation scheme outperforms exhaustive annotation for modest annotation budgets.

MasakhaNER: Named Entity Recognition for African Languages
Jade Abbott et al.

We take a step towards addressing the under-representation of the African continent in NLP research, by bringing together different stakeholders to create the first large, publicly available, high-quality dataset for named entity recognition (NER) in ten African languages. We detail the characteristics of these languages to help researchers and practitioners better understand the challenges they pose for NER tasks. We analyze our datasets and conduct an extensive empirical evaluation of state-of-the-art methods across both supervised and transfer learning settings. Finally, we release the data, code, and models to inspire future research on African NLP.

Condenser: a Pre-training Architecture for Dense Retrieval
Luyu Gao and Jamie Callan

Pre-trained Transformer language models (LM) have become go-to text representation encoders. Prior research fine-tunes deep LMs to encode text sequences such as sentences and passages into single dense vector representations for efficient text comparison and retrieval. However, dense encoders require a lot of data and sophisticated techniques to effectively train and suffer in low data situations. This paper finds a key reason is that standard LMs' internal attention structure is not ready-to-use for dense encoders, which needs to aggregate text information into the dense representation. We propose to pre-train towards dense encoder with a novel Transformer architecture, Condenser, where LM prediction CONditions on DENSE Representation. Our experiments show Condenser improves over standard LM by large margins on various text retrieval and similarity tasks.

Coarse2Fine: Fine-grained Text Classification on Coarsely-grained Annotated Data
Dheeraj Mekala, Varun Gangal, and Jingbo Shang

Existing text classification methods mainly focus on a fixed label set, whereas many real-world applications require extending to new fine-grained classes as the number of samples per label increases. To accommodate such requirements, we introduce a new problem called coarse-to-fine grained classification, which aims to perform fine-grained classification on coarsely annotated data. Instead of asking for new fine-grained human annotations, we opt to leverage label surface names as the only human guidance and weave in rich pre-trained generative language models into the iterative weak supervision strategy. Specifically, we first propose a label-conditioned fine-tuning formulation to attune these generators for our task. Furthermore, we devise a regularization objective based on the coarse-fine label constraints derived from our problem setting, giving us even further improvements over the prior formulation. Our framework uses the fine-tuned generative models to sample pseudo-training data for training the classifier, and bootstraps on real unlabeled data for model refinement. Extensive experiments and case studies on two real-world datasets demonstrate superior performance over SOTA zero-shot classification baselines.

Text2Mol: Cross-Modal Molecule Retrieval with Natural Language Queries
Carl Edwards, ChengXiang Zhai, and Heng Ji

We propose a new task, Text2Mol, to retrieve molecules using natural language descriptions as queries. Natural language and molecules encode information in very different ways, which leads to the exciting but challenging problem of integrating these two very different modalities. Although some work has been done on text-based retrieval and structure-based retrieval, this new task requires integrating molecules and natural language more directly. Moreover, this can be viewed as an especially challenging cross-lingual retrieval problem by considering the molecules as a language with a very unique grammar. We construct a paired dataset of molecules and their corresponding text descriptions, which we use to learn an aligned common semantic embedding space for retrieval. We extend this to create a cross-modal attention-based model for explainability and reranking by interpreting the attentions as association rules. We also employ an ensemble approach to integrate our different architectures, which significantly improves results from 0.372 to 0.499 MRR. This new multimodal approach opens a new perspective on solving problems in chemistry literature understanding and molecular machine learning.

Discretized Integrated Gradients for Explaining Language Models
Soumya Sanyal and Xiang Ren

As a prominent attribution-based explanation algorithm, Integrated Gradients (IG) is widely adopted due to its desirable explanation axioms and the ease of gradient computation. It measures feature importance by averaging the model's output gradient interpolated along a straight-line path in the input data space. However, such straight-line interpolated points are not representative of text data due to the inherent discreteness of the word embedding space. This questions the faithfulness of the gradients computed at the interpolated points and consequently, the quality of the generated explanations. Here we propose Discretized Integrated Gradients (DIG), which allows effective attribution along non-linear interpolation paths. We develop two interpolation

strategies for the discrete word embedding space that generates interpolation points that lie close to actual words in the embedding space, yielding more faithful gradient computation. We demonstrate the effectiveness of DIG over IG through experimental and human evaluations on multiple sentiment classification datasets. We provide the source code of DIG to encourage reproducible research.

Disentangling Representations of Text by Masking Transformers
Xiongyi Zhang, Jan-Willem van de Meent, and Byron Wallace

Representations from large pretrained models such as BERT encode a range of features into monolithic vectors, affording strong predictive accuracy across a range of downstream tasks. In this paper we explore whether it is possible to learn disentangled representations by identifying existing subnetworks within pretrained models that encode distinct, complementary aspects. Concretely, we learn binary masks over transformer weights or hidden units to uncover subsets of features that correlate with a specific factor of variation; this eliminates the need to train a disentangled model from scratch for a particular task. We evaluate this method with respect to its ability to disentangle representations of sentiment from genre in movie reviews, toxicity from dialect in Tweets, and syntax from semantics. By combining masking with magnitude pruning we find that we can identify sparse subnetworks within BERT that strongly encode particular aspects (e.g., semantics) while only weakly encoding others (e.g., syntax). Moreover, despite only learning masks, disentanglement-via-masking performs as well as — and often better than —previously proposed methods based on variational autoencoders and adversarial training.

Rationales for Sequential Predictions
Keyon Vafa et al.

Sequence models are a critical component of modern NLP systems, but their predictions are difficult to explain. We consider model explanations though rationales, subsets of context that can explain individual model predictions. We find sequential rationales by solving a combinatorial optimization: the best rationale is the smallest subset of input tokens that would predict the same output as the full sequence. Enumerating all subsets is intractable, so we propose an efficient greedy algorithm to approximate this objective. The algorithm, which is called greedy rationalization, applies to any model. For this approach to be effective, the model should form compatible conditional distributions when making predictions on incomplete subsets of the context. This condition can be enforced with a short fine-tuning step. We study greedy rationalization on language modeling and machine translation. Compared to existing baselines, greedy rationalization is best at optimizing the sequential objective and provides the most faithful rationales. On a new dataset of annotated sequential rationales, greedy rationales are most similar to human rationales.

All Bark and No Bite: Rogue Dimensions in Transformer Language Models Obscure Representational Quality
William Timkey and Marten van Schijndel

Similarity measures are a vital tool for understanding how language models represent and process language. Standard representational similarity measures such as cosine similarity and Euclidean distance have been successfully used in static word embedding models to understand how words cluster in semantic space. Recently, these measures have been applied to embeddings from contextualized models such as BERT and GPT-2. In this work, we call into question the informativity of such measures for contextualized language models. We find that a small number of rogue dimensions, often just 1-3, dominate these measures. Moreover, we find a striking mismatch between the dimensions that dominate similarity measures and those which are important to the behavior of the model. We show that simple postprocessing techniques such as standardization are able to correct for rogue dimensions and reveal underlying representational quality. We argue that accounting for rogue dimensions is essential for any similarity-based analysis of contextual language models.

Do Long-Range Language Models Actually Use Long-Range Context?
Simeng Sun et al.

Language models are generally trained on short, truncated input sequences, which limits their ability to use discourse-level information present in long-range context to improve their predictions. Recent efforts to improve the efficiency of self-attention have led to a proliferation of long-range Transformer language models, which can process much longer sequences than models of the past. However, the ways in which such models take advantage of the long-range context remain unclear. In this paper, we perform a fine-grained analysis of two long-range Transformer language models (including the Routing Transformer, which achieves state-of-the-art perplexity on the PG-19 long-sequence LM benchmark dataset) that accept input sequences of up to 8K tokens. Our results reveal that providing long-range context (i.e., beyond the previous 2K tokens) to these models only improves their predictions on a small set of tokens (e.g., those that can be copied from the distant context) and does not help at all for sentence-level prediction tasks. Finally, we discover that PG-19 contains a variety of different document types and domains, and that long-range context helps most for literary novels (as opposed to textbooks or magazines).

FastIF: Scalable Influence Functions for Efficient Model Interpretation and Debugging
Han Guo et al.

Influence functions approximate the "influences" of training data-points for test predictions and have a wide variety of applications. Despite the popularity, their computational cost does not scale well with model and training data size. We present FastIF, a set of simple modifications to influence functions that significantly im-

proves their run-time. We use k-Nearest Neighbors (kNN) to narrow the search space down to a subset of good candidate data points, identify the configurations that best balance the speed-quality trade-off in estimating the inverse Hessian-vector product, and introduce a fast parallel variant. Our proposed method achieves about 80X speedup while being highly correlated with the original influence values. With the availability of the fast influence functions, we demonstrate their usefulness in four applications. First, we examine whether influential data-points can "explain" test time behavior using the framework of simulatability. Second, we visualize the influence interactions between training and test data-points. Third, we show that we can correct model errors by additional fine-tuning on certain influential data-points, improving the accuracy of a trained MultiNLI model by 2.5% on the HANS dataset. Finally, we experiment with a similar setup but fine-tuning on datapoints not seen during training, improving the model accuracy by 2.8% and 1.7% on HANS and ANLI datasets respectively. Overall, our fast influence functions can be efficiently applied to large models and datasets, and our experiments demonstrate the potential of influence functions in model interpretation and correcting model errors.

The World of an Octopus: How Reporting Bias Influences a Language Model's Perception of Color

Cory Paik et al.

Recent work has raised concerns about the inherent limitations of text-only pretraining. In this paper, we first demonstrate that reporting bias, the tendency of people to not state the obvious, is one of the causes of this limitation, and then investigate to what extent multimodal training can mitigate this issue. To accomplish this, we 1) generate the Color Dataset (CoDa), a dataset of human-perceived color distributions for 521 common objects; 2) use CoDa to analyze and compare the color distribution found in text, the distribution captured by language models, and a human's perception of color; and 3) investigate the performance differences between text-only and multimodal models on CoDa. Our results show that the distribution of colors that a language model recovers correlates more strongly with the inaccurate distribution found in text than with the ground-truth, supporting the claim that reporting bias negatively impacts and inherently limits text-only training. We then demonstrate that multimodal models can leverage their visual training to mitigate these effects, providing a promising avenue for future research.

Studying word order through iterative shuffling

Nikolay Malkin et al.

As neural language models approach human performance on NLP benchmark tasks, their advances are widely seen as evidence of an increasingly complex understanding of syntax. This view rests upon a hypothesis that has not yet been empirically tested: that word order encodes meaning essential to performing these tasks. We refute this hypothesis in many cases: in the GLUE suite and in various genres of English text, the words in a sentence or phrase can rarely be permuted to form a phrase carrying substantially different information. Our surprising result relies on inference by iterative shuffling (IBIS), a novel, efficient procedure that finds the ordering of a bag of words having the highest likelihood under a fixed language model. IBIS can use any black-box model without additional training and is superior to existing word ordering algorithms. Coalescing our findings, we discuss how shuffling inference procedures such as IBIS can benefit language modeling and constrained generation.

An Information-Theoretic Characterization of Morphological Fusion

Neil Rathi, Michael Hahn, and Richard Futrell

Linguistic typology generally divides synthetic languages into groups based on their morphological fusion. However, this measure has long been thought to be best considered a matter of degree. We present an information-theoretic measure, called informational fusion, to quantify the degree of fusion of a given set of morphological features in a surface form, which naturally provides such a graded scale. Informational fusion is able to encapsulate not only concatenative, but also nonconcatenative morphological systems (e.g. Arabic), abstracting away from any notions of morpheme segmentation. We then show, on a sample of twenty-one languages, that our measure recapitulates the usual linguistic classifications for concatenative systems, and provides new measures for nonconcatenative ones. We also evaluate the long-standing hypotheses that more frequent forms are more fusional, and that paradigm size anticorrelates with degree of fusion. We do not find evidence for the idea that languages have characteristic levels of fusion; rather, the degree of fusion varies across part-of-speech within languages.

Frequency Effects on Syntactic Rule Learning in Transformers

Jason Wei et al.

Pre-trained language models perform well on a variety of linguistic tasks that require symbolic reasoning, raising the question of whether such models implicitly represent abstract symbols and rules. We investigate this question using the case study of BERT's performance on English subject–verb agreement. Unlike prior work, we train multiple instances of BERT from scratch, allowing us to perform a series of controlled interventions at pre-training time. We show that BERT often generalizes well to subject–verb pairs that never occurred in training, suggesting a degree of rule-governed behavior. We also find, however, that performance is heavily influenced by word frequency, with experiments showing that both the absolute frequency of a verb form, as well as the frequency relative to the alternate inflection, are causally implicated in the predictions BERT makes at inference time. Closer analysis of these frequency effects reveals that BERT's behavior is consistent with a system that correctly applies the SVA rule in general but struggles to overcome strong training priors and to estimate agreement features (singular vs. plural) on infrequent lexical items.

A surprisal–duration trade-off across and within the world's languages
Tiago Pimentel et al.

While there exist scores of natural languages, each with its unique features and idiosyncrasies, they all share a unifying theme: enabling human communication. We may thus reasonably predict that human cognition shapes how these languages evolve and are used. Assuming that the capacity to process information is roughly constant across human populations, we expect a surprisal–duration trade-off to arise both across and within languages. We analyse this trade-off using a corpus of 600 languages and, after controlling for several potential confounds, we find strong supporting evidence in both settings. Specifically, we find that, on average, phones are produced faster in languages where they are less surprising, and vice versa. Further, we confirm that more surprising phones are longer, on average, in 319 languages out of the 600. We thus conclude that there is strong evidence of a surprisal–duration trade-off in operation, both across and within the world's languages.

On the Relationships Between the Grammatical Genders of Inanimate Nouns and Their Co-Occurring Adjectives and Verbs
Adina Williams et al.

We use large-scale corpora in six different gendered languages, along with tools from NLP and information theory, to test whether there is a relationship between the grammatical genders of inanimate nouns and the adjectives used to describe those nouns. For all six languages, we find that there is a statistically significant relationship. We also find that there are statistically significant relationships between the grammatical genders of inanimate nouns and the verbs that take those nouns as direct objects, as indirect objects, and as subjects. We defer a deeper investigation of these relationships for future work.

Editing Factual Knowledge in Language Models
Nicola De Cao, Wilker Aziz, and Ivan Titov

The factual knowledge acquired during pre-training and stored in the parameters of Language Models (LMs) can be useful in downstream tasks (e.g., question answering or textual inference). However, some facts can be incorrectly induced or become obsolete over time. We present KnowledgeEditor, a method which can be used to edit this knowledge and, thus, fix 'bugs' or unexpected predictions without the need for expensive re-training or fine-tuning. Besides being computationally efficient, KnowledgeEditordoes not require any modifications in LM pre-training (e.g., the use of meta-learning). In our approach, we train a hyper-network with constrained optimization to modify a fact without affecting the rest of the knowledge; the trained hyper-network is then used to predict the weight update at test time. We show KnowledgeEditor's efficacy with two popular architectures and knowledge-intensive tasks: i) a BERT model fine-tuned for fact-checking, and ii) a sequence-to-sequence BART model for question answering. With our method, changing a prediction on the specific wording of a query tends to result in a consistent change in predictions also for its paraphrases. We show that this can be further encouraged by exploiting (e.g., automatically-generated) paraphrases during training. Interestingly, our hyper-network can be regarded as a 'probe' revealing which components need to be changed to manipulate factual knowledge; our analysis shows that the updates tend to be concentrated on a small subset of components. Source code available at https://github.com/nicola-decao/KnowledgeEditor

Pushing on Text Readability Assessment: A Transformer Meets Handcrafted Linguistic Features
Bruce W. Lee, Yoo Sung Jang, and Jason Lee

We report two essential improvements in readability assessment: 1. three novel features in advanced semantics and 2. the timely evidence that traditional ML models (e.g. Random Forest, using handcrafted features) can combine with transformers (e.g. RoBERTa) to augment model performance. First, we explore suitable transformers and traditional ML models. Then, we extract 255 handcrafted linguistic features using self-developed extraction software. Finally, we assemble those to create several hybrid models, achieving state-of-the-art (SOTA) accuracy on popular datasets in readability assessment. The use of handcrafted features help model performance on smaller datasets. Notably, our RoBERTA-RF-T1 hybrid achieves the near-perfect classification accuracy of 99%, a 20.3% increase from the previous SOTA.

Relational World Knowledge Representation in Contextual Language Models: A Review
Tara Safavi and Danai Koutra

Relational knowledge bases (KBs) are commonly used to represent world knowledge in machines. However, while advantageous for their high degree of precision and interpretability, KBs are usually organized according to manually-defined schemas, which limit their expressiveness and require significant human efforts to engineer and maintain. In this review, we take a natural language processing perspective to these limitations, examining how they may be addressed in part by training deep contextual language models (LMs) to internalize and express relational knowledge in more flexible forms. We propose to organize knowledge representation strategies in LMs by the level of KB supervision provided, from no KB supervision at all to entity- and relation-level supervision. Our contributions are threefold: (1) We provide a high-level, extensible taxonomy for knowledge representation in LMs; (2) Within our taxonomy, we highlight notable models, evaluation tasks, and findings, in order to provide an up-to-date review of current knowledge representation capabilities in LMs; and (3) We suggest future research directions that build upon the complementary aspects of LMs and KBs as knowledge representations.

Certified Robustness to Programmable Transformations in LSTMs
Yuhao Zhang, Aws Albarghouthi, and Loris D'Antoni

Deep neural networks for natural language processing are fragile in the face of adversarial examples—small input perturbations, like synonym substitution or word duplication, which cause a neural network to change its prediction. We present an approach to certifying the robustness of LSTMs (and extensions of LSTMs) and training models that can be efficiently certified. Our approach can certify robustness to intractably large perturbation spaces defined programmatically in a language of string transformations. Our evaluation shows that (1) our approach can train models that are more robust to combinations of string transformations than those produced using existing techniques; (2) our approach can show high certification accuracy of the resulting models.

Types of Out-of-Distribution Texts and How to Detect Them
Udit Arora, William Huang, and He He

Despite agreement on the importance of detecting out-of-distribution (OOD) examples, there is little consensus on the formal definition of the distribution shifts of OOD examples and how to best detect them. We categorize these examples as exhibiting a background shift or semantic shift, and find that the two major approaches to OOD detection, calibration and density estimation (language modeling for text), have distinct behavior on these types of OOD data. Across 14 pairs of in-distribution and OOD English natural language understanding datasets, we find that density estimation methods consistently beat calibration methods in background shift settings and perform worse in semantic shift settings. In addition, we find that both methods generally fail to detect examples from challenge data, indicating that these examples constitute a different type of OOD data. Overall, while the categorization we apply explains many of the differences between the two methods, our results call for a more explicit definition of OOD to create better benchmarks and build detectors that can target the type of OOD data expected at test time.

Value-aware Approximate Attention
Ankit Gupta and Jonathan Berant

Following the success of dot-product attention in Transformers, numerous approximations have been recently proposed to address its quadratic complexity with respect to the input length. However, all approximations thus far have ignored the contribution of the *value vectors* to the quality of approximation. In this work, we argue that research efforts should be directed towards approximating the true output of the attention sub-layer, which includes the value vectors. We propose a value-aware objective, and show theoretically and empirically that an optimal approximation of a value-aware objective substantially outperforms an optimal approximation that ignores values, in the context of language modeling. Moreover, we show that the choice of kernel function for computing attention similarity can substantially affect the quality of sparse approximations, where kernel functions that are less skewed are more affected by the value vectors.

Effects of Parameter Norm Growth During Transformer Training: Inductive Bias from Gradient Descent
William Merrill et al.

The capacity of neural networks like the widely adopted transformer is known to be very high. Evidence is emerging that they learn successfully due to inductive bias in the training routine, typically a variant of gradient descent (GD). To better understand this bias, we study the tendency for transformer parameters to grow in magnitude (ℓ_2 norm) during training, and its implications for the emergent representations within self attention layers. Empirically, we document norm growth in the training of transformer language models, including T5 during its pretraining. As the parameters grow in magnitude, we prove that the network approximates a discretized network with saturated activation functions. Such "saturated" networks are known to have a reduced capacity compared to the full network family that can be described in terms of formal languages and automata. Our results suggest saturation is a new characterization of an inductive bias implicit in GD of particular interest for NLP. We leverage the emergent discrete structure in a saturated transformer to analyze the role of different attention heads, finding that some focus locally on a small number of positions, while other heads compute global averages, allowing counting. We believe understanding the interplay between these two capabilities may shed further light on the structure of computation within large transformers.

Self-training with Few-shot Rationalization
Meghana Moorthy Bhat, Alessandro Sordoni, and Subhabrata Mukherjee

While pre-trained language models have obtained state-of-the-art performance for several natural language understanding tasks, they are quite opaque in terms of their decision-making process. While some recent works focus on rationalizing neural predictions by highlighting salient concepts in the text as justifications or rationales, they rely on thousands of labeled training examples for both task labels as well as annotated rationales for every instance. Such extensive large-scale annotations are infeasible to obtain for many tasks. To this end, we develop a multi-task teacher-student framework based on self-training pre-trained language models with limited task-specific labels and rationales and judicious sample selection to learn from informative pseudo-labeled examples. We study several characteristics of what constitutes a good rationale and demonstrate that the neural model performance can be significantly improved by making it aware of its rationalized predictions, particularly in low-resource settings. Extensive experiments in several benchmark datasets demonstrate the effectiveness of our approach.

Adversarial Regularization as Stackelberg Game: An Unrolled Optimization Approach
Simiao Zuo et al.

Adversarial regularization has been shown to improve the generalization performance of deep learning models in various natural language processing tasks. Existing works usually formulate the method as a zero-sum game, which is solved by alternating gradient descent/ascent algorithms. Such a formulation treats the adversarial and the defending players equally, which is undesirable because only the defending player contributes to the generalization performance. To address this issue, we propose Stackelberg Adversarial Regularization (SALT), which formulates adversarial regularization as a Stackelberg game. This formulation induces a competition between a leader and a follower, where the follower generates perturbations, and the leader trains the model subject to the perturbations. Different from conventional approaches, in SALT, the leader is in an advantageous position. When the leader moves, it recognizes the strategy of the follower and takes the anticipated follower's outcomes into consideration. Such a leader's advantage enables us to improve the model fitting to the unperturbed data. The leader's strategic information is captured by the Stackelberg gradient, which is obtained using an unrolling algorithm. Our experimental results on a set of machine translation and natural language understanding tasks show that SALT outperforms existing adversarial regularization baselines across all tasks. Our code is publicly available.

Foreseeing the Benefits of Incidental Supervision
Hangfeng He et al.

Real-world applications often require improved models by leveraging *a range of cheap incidental supervision signals*. These could include partial labels, noisy labels, knowledge-based constraints, and cross-domain or cross-task annotations – all having statistical associations with gold annotations but not exactly the same. However, we currently lack a principled way to measure the benefits of these signals to a given target task, and the common practice of evaluating these benefits is through exhaustive experiments with various models and hyperparameters. This paper studies whether we can, *in a single framework, quantify the benefits of various types of incidental signals for a given target task without going through combinatorial experiments*. We propose a unified PAC-Bayesian motivated informativeness measure, PABI, that characterizes the uncertainty reduction provided by incidental supervision signals. We demonstrate PABI's effectiveness by quantifying the value added by various types of incidental signals to sequence tagging tasks. Experiments on named entity recognition (NER) and question answering (QA) show that PABI's predictions correlate well with learning performance, providing a promising way to determine, ahead of learning, which supervision signals would be beneficial.

FAME: Feature-Based Adversarial Meta-Embeddings for Robust Input Representations
Lukas Lange et al.

Combining several embeddings typically improves performance in downstream tasks as different embeddings encode different information. It has been shown that even models using embeddings from transformers still benefit from the inclusion of standard word embeddings. However, the combination of embeddings of different types and dimensions is challenging. As an alternative to attention-based meta-embeddings, we propose feature-based adversarial meta-embeddings (FAME) with an attention function that is guided by features reflecting word-specific properties, such as shape and frequency, and show that this is beneficial to handle subword-based embeddings. In addition, FAME uses adversarial training to optimize the mappings of differently-sized embeddings to the same space. We demonstrate that FAME works effectively across languages and domains for sequence labeling and sentence classification, in particular in low-resource settings. FAME sets the new state of the art for POS tagging in 27 languages, various NER settings and question classification in different domains.

ReGen: Reinforcement Learning for Text and Knowledge Base Generation using Pretrained Language Models
Pierre Dognin et al.

Automatic construction of relevant Knowledge Bases (KBs) from text, and generation of semantically meaningful text from KBs are both long-standing goals in Machine Learning. In this paper, we present ReGen, a bidirectional generation of text and graph leveraging Reinforcement Learning to improve performance. Graph linearization enables us to re-frame both tasks as a sequence to sequence generation problem regardless of the generative direction, which in turn allows the use of Reinforcement Learning for sequence training where the model itself is employed as its own critic leading to Self-Critical Sequence Training (SCST). We present an extensive investigation demonstrating that the use of RL via SCST benefits graph and text generation on WebNLG+ 2020 and TekGen datasets. Our system provides state-of-the-art results on WebNLG+ 2020 by significantly improving upon published results from the WebNLG 2020+ Challenge for both text-to-graph and graph-to-text generation tasks. More details at https://github.com/IBM/regen.

Sentence Bottleneck Autoencoders from Transformer Language Models
Ivan Montero, Nikolaos Pappas, and Noah A. Smith

Representation learning for text via pretraining a language model on a large corpus has become a standard starting point for building NLP systems. This approach stands in contrast to autoencoders, also trained on raw text, but with the objective of learning to encode each input as a vector that allows full reconstruction. Autoencoders are attractive because of their latent space structure and generative properties. We therefore explore the construction of a sentence-level autoencoder from a pretrained, frozen transformer language model. We adapt the masked language modeling objective as a generative, denoising one, while only training a sentence bottleneck and a single-layer modified transformer decoder. We demonstrate that the sentence representations discovered by our model achieve better quality than previous methods that extract representations from pretrained

transformers on text similarity tasks, style transfer (an example of controlled generation), and single-sentence classification tasks in the GLUE benchmark, while using fewer parameters than large pretrained models.

Contrastive Out-of-Distribution Detection for Pretrained Transformers
Wenxuan Zhou, Fangyu Liu, and Muhao Chen

Pretrained Transformers achieve remarkable performance when training and test data are from the same distribution. However, in real-world scenarios, the model often faces out-of-distribution (OOD) instances that can cause severe semantic shift problems at inference time. Therefore, in practice, a reliable model should identify such instances, and then either reject them during inference or pass them over to models that handle another distribution. In this paper, we develop an unsupervised OOD detection method, in which only the in-distribution (ID) data are used in training. We propose to fine-tune the Transformers with a contrastive loss, which improves the compactness of representations, such that OOD instances can be better differentiated from ID ones. These OOD instances can then be accurately detected using the Mahalanobis distance in the model's penultimate layer. We experiment with comprehensive settings and achieve near-perfect OOD detection performance, outperforming baselines drastically. We further investigate the rationales behind the improvement, finding that more compact representations through margin-based contrastive learning bring the improvement. We release our code to the community for future research.

Competency Problems: On Finding and Removing Artifacts in Language Data
Matt Gardner et al.

Much recent work in NLP has documented dataset artifacts, bias, and spurious correlations between input features and output labels. However, how to tell which features have "spurious" instead of legitimate correlations is typically left unspecified. In this work we argue that for complex language understanding tasks, all simple feature correlations are spurious, and we formalize this notion into a class of problems which we call competency problems. For example, the word "amazing" on its own should not give information about a sentiment label independent of the context in which it appears, which could include negation, metaphor, sarcasm, etc. We theoretically analyze the difficulty of creating data for competency problems when human bias is taken into account, showing that realistic datasets will increasingly deviate from competency problems as dataset size increases. This analysis gives us a simple statistical test for dataset artifacts, which we use to show more subtle biases than were described in prior work, including demonstrating that models are inappropriately affected by these less extreme biases. Our theoretical treatment of this problem also allows us to analyze proposed solutions, such as making local edits to dataset instances, and to give recommendations for future data collection and model design efforts that target competency problems.

Softmax Tree: An Accurate, Fast Classifier When the Number of Classes Is Large
Arman Zharmagambetov, Magzhan Gabidolla, and Miguel A. Carreira-Perpinan

Classification problems having thousands or more classes naturally occur in NLP, for example language models or document classification. A softmax or one-vs-all classifier naturally handles many classes, but it is very slow at inference time, because every class score must be calculated to find the top class. We propose the "softmax tree", consisting of a binary tree having sparse hyperplanes at the decision nodes (which make hard, not soft, decisions) and small softmax classifiers at the leaves. This is much faster at inference because the input instance follows a single path to a leaf (whose length is logarithmic on the number of leaves) and the softmax classifier at each leaf operates on a small subset of the classes. Although learning accurate tree-based models has proven difficult in the past, we are able to overcome this by using a variation of a recent algorithm, tree alternating optimization (TAO). Compared to a softmax and other classifiers, the resulting softmax trees are both more accurate in prediction and faster in inference, as shown in NLP problems having from one thousand to one hundred thousand classes.

Detecting Local Insights from Global Labels: Supervised & Zero-Shot Sequence Labeling via a Convolutional Decomposition
Allen Schmaltz

We propose a new, more actionable view of neural network interpretability and data analysis by leveraging the remarkable matching effectiveness of representations derived from deep networks, guided by an approach for class-conditional feature detection. The decomposition of the filter-n-gram interactions of a convolutional neural network (CNN) and a linear layer over a pre-trained deep network yields a strong binary sequence labeler, with flexibility in producing predictions at—and defining loss functions for—varying label granularities, from the fully supervised sequence labeling setting to the challenging zero-shot sequence labeling setting, in which we seek token-level predictions but only have document-level labels for training. From this sequence-labeling layer we derive dense representations of the input that can then be matched to instances from training, or a support set with known labels. Such introspection with inference-time decision rules provides a means, in some settings, of making local updates to the model by altering the labels or instances in the support set without re-training the full model. Finally, we construct a particular K-nearest neighbors (K-NN) model from matched exemplar representations that approximates the original model's predictions and is at least as effective a predictor with respect to the ground-truth labels. This additionally yields interpretable heuristics at the token level for determining when predictions are less likely to be reliable, and for screening input dissimilar to the support set. In effect, we show that we can transform the deep network into a simple weighting over exemplars and associated labels, yielding an introspectable—and modestly updatable—version of the original model.

Sensitivity as a Complexity Measure for Sequence Classification Tasks
Michael Hahn, Dan Jurafsky, and Richard Futrell

We introduce a theoretical framework for understanding and predicting the complexity of sequence classification tasks, using a novel extension of the theory of Boolean function sensitivity. The sensitivity of a function, given a distribution over input sequences, quantifies the number of disjoint subsets of the input sequence that can each be individually changed to change the output. We argue that standard sequence classification methods are biased towards learning low-sensitivity functions, so that tasks requiring high sensitivity are more difficult. To that end, we show analytically that simple lexical classifiers can only express functions of bounded sensitivity, and we show empirically that low-sensitivity functions are easier to learn for LSTMs. We then estimate sensitivity on 15 NLP tasks, finding that sensitivity is higher on challenging tasks collected in GLUE than on simple text classification tasks, and that sensitivity predicts the performance both of simple lexical classifiers and of vanilla BiLSTMs without pretrained contextualized embeddings. Within a task, sensitivity predicts which inputs are hard for such simple models. Our results suggest that the success of massively pretrained contextual representations stems in part because they provide representations from which information can be extracted by low-sensitivity decoders.

GFST: Gender-Filtered Self-Training for More Accurate Gender in Translation
Prafulla Kumar Choubey et al.

Targeted evaluations have found that machine translation systems often output incorrect gender in translations, even when the gender is clear from context. Furthermore, these incorrectly gendered translations have the potential to reflect or amplify social biases. We propose gender-filtered self-training (GFST) to improve gender translation accuracy on unambiguously gendered inputs. Our GFST approach uses a source monolingual corpus and an initial model to generate gender-specific pseudo-parallel corpora which are then filtered and added to the training data. We evaluate GFST on translation from English into five languages, finding that it improves gender accuracy without damaging generic quality. We also show the viability of GFST on several experimental settings, including re-training from scratch, fine-tuning, controlling the gender balance of the data, forward translation, and back-translation.

Robust Open-Vocabulary Translation from Visual Text Representations
Elizabeth Salesky, David Etter, and Matt Post

Machine translation models have discrete vocabularies and commonly use subword segmentation techniques to achieve an 'open vocabulary.' This approach relies on consistent and correct underlying unicode sequences, and makes models susceptible to degradation from common types of noise and variation. Motivated by the robustness of human language processing, we propose the use of visual text representations, which dispense with a finite set of text embeddings in favor of continuous vocabularies created by processing visually rendered text with sliding windows. We show that models using visual text representations approach or match performance of traditional text models on small and larger datasets. More importantly, models with visual embeddings demonstrate significant robustness to varied types of noise, achieving e.g., 25.9 BLEU on a character permuted German–English task where subword models degrade to 1.9.

"Wikily" Supervised Neural Translation Tailored to Cross-Lingual Tasks
Mohammad Sadegh Rasooli, Chris Callison-Burch, and Derry Tanti Wijaya

We present a simple but effective approach for leveraging Wikipedia for neural machine translation as well as cross-lingual tasks of image captioning and dependency parsing without using any direct supervision from external parallel data or supervised models in the target language. We show that first sentences and titles of linked Wikipedia pages, as well as cross-lingual image captions, are strong signals for a seed parallel data to extract bilingual dictionaries and cross-lingual word embeddings for mining parallel text from Wikipedia. Our final model achieves high BLEU scores that are close to or sometimes higher than strong *supervised* baselines in low-resource languages; e.g. supervised BLEU of 4.0 versus 12.1 from our model in English-to-Kazakh. Moreover, we tailor our *wikily* translation models to unsupervised image captioning, and cross-lingual dependency parser transfer. In image captioning, we train a multi-tasking machine translation and image captioning pipeline for Arabic and English from which the Arabic training data is a *wikily* translation of the English captioning data. Our captioning results on Arabic are slightly *better* than that of its supervised model. In dependency parsing, we translate a large amount of monolingual text, and use it as an artificial training data in an *annotation projection* framework. We show that our model outperforms recent work on cross-lingual transfer of dependency parsers.

Don't Go Far Off: An Empirical Study on Neural Poetry Translation
Tuhin Chakrabarty, Arkadiy Saakyan, and Smaranda Muresan

Despite constant improvements in machine translation quality, automatic poetry translation remains a challenging problem due to the lack of open-sourced parallel poetic corpora, and to the intrinsic complexities involved in preserving the semantics, style and figurative nature of poetry. We present an empirical investigation for poetry translation along several dimensions: 1) size and style of training data (poetic vs. non-poetic), including a zero-shot setup; 2) bilingual vs. multilingual learning; and 3) language-family-specific models vs. mixed-language-family models. To accomplish this, we contribute a parallel dataset of poetry translations for several language pairs. Our results show that multilingual fine-tuning on poetic text significantly outperforms multilingual fine-tuning on non-poetic text that is 35X larger in size, both in terms of automatic metrics (BLEU, BERTScore, COMET) and human evaluation metrics such as faithfulness (meaning and poetic style). Moreover, multilingual fine-tuning on poetic data outperforms bilingual fine-tuning on poetic data.

Improving Zero-Shot Cross-Lingual Transfer Learning via Robust Training
Kuan-Hao Huang et al.

Pre-trained multilingual language encoders, such as multilingual BERT and XLM-R, show great potential for zero-shot cross-lingual transfer. However, these multilingual encoders do not precisely align words and phrases across languages. Especially, learning alignments in the multilingual embedding space usually requires sentence-level or word-level parallel corpora, which are expensive to be obtained for low-resource languages. An alternative is to make the multilingual encoders more robust; when fine-tuning the encoder using downstream task, we train the encoder to tolerate noise in the contextual embedding spaces such that even if the representations of different languages are not aligned well, the model can still achieve good performance on zero-shot cross-lingual transfer. In this work, we propose a learning strategy for training robust models by drawing connections between adversarial examples and the failure cases of zero-shot cross-lingual transfer. We adopt two widely used robust training methods, adversarial training and randomized smoothing, to train the desired robust model. The experimental results demonstrate that robust training improves zero-shot cross-lingual transfer on text classification tasks. The improvement is more significant in the generalized cross-lingual transfer setting, where the pair of input sentences belong to two different languages.

BERT, mBERT, or BiBERT? A Study on Contextualized Embeddings for Neural Machine Translation
Haoran Xu, Benjamin Van Durme, and Kenton Murray

The success of bidirectional encoders using masked language models, such as BERT, on numerous natural language processing tasks has prompted researchers to attempt to incorporate these pre-trained models into neural machine translation (NMT) systems. However, proposed methods for incorporating pre-trained models are non-trivial and mainly focus on BERT, which lacks a comparison of the impact that other pre-trained models may have on translation performance. In this paper, we demonstrate that simply using the output (contextualized embeddings) of a tailored and suitable bilingual pre-trained language model (dubbed BiBERT) as the input of the NMT encoder achieves state-of-the-art translation performance. Moreover, we also propose a stochastic layer selection approach and a concept of a dual-directional translation model to ensure the sufficient utilization of contextualized embeddings. In the case of without using back translation, our best models achieve BLEU scores of 30.45 for En→De and 38.61 for De→En on the IWSLT'14 dataset, and 31.26 for En→De and 34.94 for De→En on the WMT'14 dataset, which exceeds all published numbers.

One Source, Two Targets: Challenges and Rewards of Dual Decoding
Jitao Xu and François Yvon

Machine translation is generally understood as generating one target text from an input source document. In this paper, we consider a stronger requirement: to jointly generate two texts so that each output side effectively depends on the other. As we discuss, such a device serves several practical purposes, from multi-target machine translation to the generation of controlled variations of the target text. We present an analysis of possible implementations of dual decoding, and experiment with four applications. Viewing the problem from multiple angles allows us to better highlight the challenges of dual decoding and to also thoroughly analyze the benefits of generating matched, rather than independent, translations.

Boosting Cross-Lingual Transfer via Self-Learning with Uncertainty Estimation
Liyan Xu et al.

Recent multilingual pre-trained language models have achieved remarkable zero-shot performance, where the model is only finetuned on one source language and directly evaluated on target languages. In this work, we propose a self-learning framework that further utilizes unlabeled data of target languages, combined with uncertainty estimation in the process to select high-quality silver labels. Three different uncertainties are adapted and analyzed specifically for the cross lingual transfer: Language Heteroscedastic/Homoscedastic Uncertainty (LEU/LOU), Evidential Uncertainty (EVI). We evaluate our framework with uncertainties on two cross-lingual tasks including Named Entity Recognition (NER) and Natural Language Inference (NLI) covering 40 languages in total, which outperforms the baselines significantly by 10 F1 for NER on average and 2.5 accuracy for NLI.

Cross-Attention is All You Need: Adapting Pretrained Transformers for Machine Translation
Mozhdeh Gheini, Xiang Ren, and Jonathan May

We study the power of cross-attention in the Transformer architecture within the context of transfer learning for machine translation, and extend the findings of studies into cross-attention when training from scratch. We conduct a series of experiments through fine-tuning a translation model on data where either the source or target language has changed. These experiments reveal that fine-tuning only the cross-attention parameters is nearly as effective as fine-tuning all parameters (i.e., the entire translation model). We provide insights into why this is the case and observe that limiting fine-tuning in this manner yields cross-lingually aligned embeddings. The implications of this finding for researchers and practitioners include a mitigation of catastrophic forgetting, the potential for zero-shot translation, and the ability to extend machine translation models to several new language pairs with reduced parameter storage overhead.

Levenshtein Training for Word-level Quality Estimation
Shuoyang Ding et al.

We propose a novel scheme to use the Levenshtein Transformer to perform the task of word-level quality estimation. A Levenshtein Transformer is a natural fit for this task: trained to perform decoding in an iterative manner, a Levenshtein Transformer can learn to post-edit without explicit supervision. To further minimize the mismatch between the translation task and the word-level QE task, we propose a two-stage transfer learning procedure on both augmented data and human post-editing data. We also propose heuristics to construct reference labels that are compatible with subword-level finetuning and inference. Results on WMT 2020 QE shared task dataset show that our proposed method has superior data efficiency under the data-constrained setting and competitive performance under the unconstrained setting.

Improving Multilingual Translation by Representation and Gradient Regularization
Yilin Yang et al.

Multilingual Neural Machine Translation (NMT) enables one model to serve all translation directions, including ones that are unseen during training, i.e. zero-shot translation. Despite being theoretically attractive, current models often produce low quality translations – commonly failing to even produce outputs in the right target language. In this work, we observe that off-target translation is dominant even in strong multilingual systems, trained on massive multilingual corpora. To address this issue, we propose a joint approach to regularize NMT models at both representation-level and gradient-level. At the representation level, we leverage an auxiliary target language prediction task to regularize decoder outputs to retain information about the target language. At the gradient level, we leverage a small amount of direct data (in thousands of sentence pairs) to regularize model gradients. Our results demonstrate that our approach is highly effective in both reducing off-target translation occurrences and improving zero-shot translation performance by +5.59 and +10.38 BLEU on WMT and OPUS datasets respectively. Moreover, experiments show that our method also works well when the small amount of direct data is not available.

Contrastive Conditioning for Assessing Disambiguation in MT: A Case Study of Distilled Bias
Jannis Vamvas and Rico Sennrich

Lexical disambiguation is a major challenge for machine translation systems, especially if some senses of a word are trained less often than others. Identifying patterns of overgeneralization requires evaluation methods that are both reliable and scalable. We propose contrastive conditioning as a reference-free black-box method for detecting disambiguation errors. Specifically, we score the quality of a translation by conditioning on variants of the source that provide contrastive disambiguation cues. After validating our method, we apply it in a case study to perform a targeted evaluation of sequence-level knowledge distillation. By probing word sense disambiguation and translation of gendered occupation names, we show that distillation-trained models tend to overgeneralize more than other models with a comparable BLEU score. Contrastive conditioning thus highlights a side effect of distillation that is not fully captured by standard evaluation metrics. Code and data to reproduce our findings are publicly available.

Data and Parameter Scaling Laws for Neural Machine Translation
Mitchell A Gordon, Kevin Duh, and Jared Kaplan

We observe that the development cross-entropy loss of supervised neural machine translation models scales like a power law with the amount of training data and the number of non-embedding parameters in the model. We discuss some practical implications of these results, such as predicting BLEU achieved by large scale models and predicting the ROI of labeling data in low-resource language pairs.

On the Difficulty of Translating Free-Order Case-Marking Languages
Arianna Bisazza, Ahmet Üstün, and Stephan Sportel

Identifying factors that make certain languages harder to model than others is essential to reach language equality in future Natural Language Processing technologies. Free-order case-marking languages, such as Russian, Latin or Tamil, have proved more challenging than fixed-order languages for the tasks of syntactic parsing and subject-verb agreement prediction. In this work, we investigate whether this class of languages is also more difficult to translate by state-of-the-art Neural Machine Translation models (NMT). Using a variety of synthetic languages and a newly introduced translation challenge set, we find that word order flexibility in the source language only leads to a very small loss of NMT quality, even though the core verb arguments become impossible to disambiguate in sentences without semantic cues. The latter issue is indeed solved by the addition of case marking. However, in medium- and low-resource settings, the overall NMT quality of fixed-order languages remains unmatched.

Students Who Study Together Learn Better: On the Importance of Collective Knowledge Distillation for Domain Transfer in Fact Verification
Mitch Paul Mithun, Sandeep Suntwal, and Mihai Surdeanu

While neural networks produce state-of-the- art performance in several NLP tasks, they generally depend heavily on lexicalized information, which transfer poorly between domains. Previous works have proposed delexicalization as a form of knowledge distillation to reduce the dependency on such lexical artifacts. However, a critical unsolved issue that remains is how much delexicalization to apply: a little helps reduce overfitting, but too much discards useful information. We propose Group Learning, a knowledge and model distillation approach for fact verification in which multiple student models have access to different delexicalized views of the data, but are encouraged to learn from each other through pair-wise consistency losses. In several cross-

domain experiments between the FEVER and FNC fact verification datasets, we show that our approach learns the best delexicalization strategy for the given training dataset, and outperforms state-of-the-art classifiers that rely on the original data.

Can We Improve Model Robustness through Secondary Attribute Counterfactuals?
Ananth Balashankar et al.

Developing robust NLP models that perform well on many, even small, slices of data is a significant but important challenge, with implications from fairness to general reliability. To this end, recent research has explored how models rely on spurious correlations, and how counterfactual data augmentation (CDA) can mitigate such issues. In this paper we study how and why modeling counterfactuals over multiple attributes can go significantly further in improving model performance. We propose RDI, a context-aware methodology which takes into account the impact of secondary attributes on the model's predictions and increases sensitivity for secondary attributes over reweighted counterfactually augmented data. By implementing RDI in the context of toxicity detection, we find that accounting for secondary attributes can significantly improve robustness, with improvements in sliced accuracy on the original dataset up to 7% compared to existing robustness methods. We also demonstrate that RDI generalizes to the coreference resolution task and provide guidelines to extend this to other tasks.

Semantic Novelty Detection in Natural Language Descriptions
Nianzu Ma et al.

This paper proposes to study a fine-grained semantic novelty detection task, which can be illustrated with the following example. It is normal that a person walks a dog in the park, but if someone says "A man is walking a chicken in the park", it is novel. Given a set of natural language descriptions of normal scenes, we want to identify descriptions of novel scenes. We are not aware of any existing work that solves the problem. Although existing novelty or anomaly detection algorithms are applicable, since they are usually topic-based, they perform poorly on our fine-grained semantic novelty detection task. This paper proposes an effective model (called GAT-MA) to solve the problem and also contributes a new dataset. Experimental evaluation shows that GAT-MA outperforms 11 baselines by large margins.

Jump-Starting Item Parameters for Adaptive Language Tests
Arya D. McCarthy et al.

A challenge in designing high-stakes language assessments is calibrating the test item difficulties, either a priori or from limited pilot test data. While prior work has addressed 'cold start' estimation of item difficulties without piloting, we devise a multi-task generalized linear model with BERT features to jump-start these estimates, rapidly improving their quality with as few as 500 test-takers and a small sample of item exposures (6 each) from a large item bank (4,000 items). Our joint model provides a principled way to compare test-taker proficiency, item difficulty, and language proficiency frameworks like the Common European Framework of Reference (CEFR). This also enables new item difficulty estimates without piloting them first, which in turn limits item exposure and thus enhances test item security. Finally, using operational data from the Duolingo English Test, a high-stakes English proficiency test, we find that the difficulty estimates derived using this method correlate strongly with lexico-grammatical features that correlate with reading complexity.

Long-Range Modeling of Source Code Files with eWASH: Extended Window Access by Syntax Hierarchy
Colin Clement et al.

Statistical language modeling and translation with transformers have found many successful applications in program understanding and generation tasks, setting high benchmarks for tools in modern software development environments. The finite context window of these neural models means, however, that they will be unable to leverage the entire relevant context of large files and packages for any given task. While there are many efforts to extend the context window, we introduce an architecture-independent approach for leveraging the syntactic hierarchies of source code for incorporating entire file-level context into a fixed-length window. Using concrete syntax trees of each source file we extract syntactic hierarchies and integrate them into context window by selectively removing from view more specific, less relevant scopes for a given task. We evaluate this approach on code generation tasks and joint translation of natural language and source code in Python programming language, achieving a new state-of-the-art in code completion and summarization for Python in the CodeXGLUE benchmark. We also introduce new CodeXGLUE benchmarks for user-experience-motivated tasks: code completion with normalized literals, method body completion/code summarization conditioned on file-level context.

Fast WordPiece Tokenization
Xinying Song et al.

Tokenization is a fundamental preprocessing step for almost all NLP tasks. In this paper, we propose efficient algorithms for the WordPiece tokenization used in BERT, from single-word tokenization to general text (e.g., sentence) tokenization. When tokenizing a single word, WordPiece uses a longest-match-first strategy, known as maximum matching. The best known algorithms so far are On^2 (where n is the input length) or O(nm) (where m is the maximum vocabulary token length). We propose a novel algorithm whose tokenization complexity is strictly O(n). Our method is inspired by the Aho-Corasick algorithm. We introduce additional linkages on top of the trie built from the vocabulary, allowing smart transitions when the trie matching can-

not continue. For general text, we further propose an algorithm that combines pre-tokenization (splitting the text into words) and our linear-time WordPiece method into a single pass. Experimental results show that our method is 8.2x faster than HuggingFace Tokenizers and 5.1x faster than TensorFlow Text on average for general text tokenization.

Identity-Based Patterns in Deep Convolutional Networks: Generative Adversarial Phonology and Reduplication
Gašper Beguš

This paper models unsupervised learning of an identity-based pattern (or copying) in speech called reduplication from raw continuous data with deep convolutional neural networks. We use the ciwGAN architecture Begus (2021a) in which learning of meaningful representations in speech emerges from a requirement that the CNNs generate informative data. We propose a technique to wug-test CNNs trained on speech and, based on four generative tests, argue that the network learns to represent an identity-based pattern in its latent space. By manipulating only two categorical variables in the latent space, we can actively turn an unreduplicated form into a reduplicated form with no other substantial changes to the output in the majority of cases. We also argue that the network extends the identity-based pattern to unobserved data. Exploration of how meaningful representations of identity-based patterns emerge in CNNs and how the latent space variables outside of the training range correlate with identity-based patterns in the output has general implications for neural network interpretability.

Joint Passage Ranking for Diverse Multi-Answer Retrieval
Sewon Min et al.

We study multi-answer retrieval, an under-explored problem that requires retrieving passages to cover multiple distinct answers for a given question. This task requires joint modeling of retrieved passages, as models should not repeatedly retrieve passages containing the same answer at the cost of missing a different valid answer. Prior work focusing on single-answer retrieval is limited as it cannot reason about the set of passages jointly. In this paper, we introduce JPR, a joint passage retrieval model focusing on reranking. To model the joint probability of the retrieved passages, JPR makes use of an autoregressive reranker that selects a sequence of passages, equipped with novel training and decoding algorithms. Compared to prior approaches, JPR achieves significantly better answer coverage on three multi-answer datasets. When combined with downstream question answering, the improved retrieval enables larger answer generation models since they need to consider fewer passages, establishing a new state-of-the-art.

Case-based Reasoning for Natural Language Queries over Knowledge Bases
Rajarshi Das et al.

It is often challenging to solve a complex problem from scratch, but much easier if we can access other similar problems with their solutions — a paradigm known as case-based reasoning (CBR). We propose a neuro-symbolic CBR approach (CBR-KBQA) for question answering over large knowledge bases. CBR-KBQA consists of a nonparametric memory that stores cases (question and logical forms) and a parametric model that can generate a logical form for a new question by retrieving cases that are relevant to it. On several KBQA datasets that contain complex questions, CBR-KBQA achieves competitive performance. For example, on the CWQ dataset, CBR-KBQA outperforms the current state of the art by 11% on accuracy. Furthermore, we show that CBR-KBQA is capable of using new cases *without* any further training: by incorporating a few human-labeled examples in the case memory, CBR-KBQA is able to successfully generate logical forms containing unseen KB entities as well as relations.

How much coffee was consumed during EMNLP 2019? Fermi Problems: A New Reasoning Challenge for AI
Ashwin Kalyan et al.

Many real-world problems require the combined application of multiple reasoning abilities—employing suitable abstractions, commonsense knowledge, and creative synthesis of problem-solving strategies. To help advance AI systems towards such capabilities, we propose a new reasoning challenge, namely Fermi Problems (FPs), which are questions whose answers can only be approximately estimated because their precise computation is either impractical or impossible. For example, "How much would the sea level rise if all ice in the world melted?" FPs are commonly used in quizzes and interviews to bring out and evaluate the creative reasoning abilities of humans. To do the same for AI systems, we present two datasets: 1) A collection of 1k real-world FPs sourced from quizzes and olympiads; and 2) a bank of 10k synthetic FPs of intermediate complexity to serve as a sandbox for the harder real-world challenge. In addition to question-answer pairs, the datasets contain detailed solutions in the form of an executable program and supporting facts, helping in supervision and evaluation of intermediate steps. We demonstrate that even extensively fine-tuned large-scale language models perform poorly on these datasets, on average making estimates that are off by two orders of magnitude. Our contribution is thus the crystallization of several unsolved AI problems into a single, new challenge that we hope will spur further advances in building systems that can reason.

Will this Question be Answered? Question Filtering via Answer Model Distillation for Efficient Question Answering
Siddhant Garg and Alessandro Moschitti

In this paper we propose a novel approach towards improving the efficiency of Question Answering (QA) systems by filtering out questions that will not be answered by them. This is based on an interesting new finding: the answer confidence scores of state-of-the-art QA systems can be approximated well by models solely using the input question text. This enables preemptive filtering of questions that are not answered by the system due to their answer confidence scores being lower than the system threshold. Specifically, we learn Transformer-based question models by distilling Transformer-based answering models. Our experiments on three popular QA datasets and one industrial QA benchmark demonstrate the ability of our question models to approximate the Precision/Recall curves of the target QA system well. These question models, when used as filters, can effectively trade off lower computation cost of QA systems for lower Recall, e.g., reducing computation by ~60%, while only losing ~3-4% of Recall.

Distantly-Supervised Dense Retrieval Enables Open-Domain Question Answering without Evidence Annotation

Chen Zhao et al.

Open-domain question answering answers a question based on evidence retrieved from a large corpus. State-of-the-art neural approaches require intermediate evidence annotations for training. However, such intermediate annotations are expensive, and methods that rely on them cannot transfer to the more common setting, where only question—answer pairs are available. This paper investigates whether models can learn to find evidence from a large corpus, with only distant supervision from answer labels for model training, thereby generating no additional annotation cost. We introduce a novel approach (DistDR) that iteratively improves over a weak retriever by alternately finding evidence from the up-to-date model and encouraging the model to learn the most likely evidence. Without using any evidence labels, DistDR is on par with fully-supervised state-of-the-art methods on both multi-hop and single-hop QA benchmarks. Our analysis confirms that DistDR finds more accurate evidence over iterations, which leads to model improvements. The code is available at https://github.com/henryzhao5852/DistDR.

What's in a Name? Answer Equivalence For Open-Domain Question Answering

Chenglei Si, Chen Zhao, and Jordan Boyd-Graber

A flaw in QA evaluation is that annotations often only provide one gold answer. Thus, model predictions semantically equivalent to the answer but superficially different are considered incorrect. This work explores mining alias entities from knowledge bases and using them as additional gold answers (i.e., equivalent answers). We incorporate answers for two settings: evaluation with additional answers and model training with equivalent answers. We analyse three QA benchmarks: Natural Questions, TriviaQA, and SQuAD. Answer expansion increases the exact match score on all datasets for evaluation, while incorporating it helps model training over real-world datasets. We ensure the additional answers are valid through a human post hoc evaluation.

Learning with Instance Bundles for Reading Comprehension

Dheeru Dua et al.

When training most modern reading comprehension models, all the questions associated with a context are treated as being independent from each other. However, closely related questions and their corresponding answers are not independent, and leveraging these relationships could provide a strong supervision signal to a model. Drawing on ideas from contrastive estimation, we introduce several new supervision losses that compare question-answer scores across multiple related instances. Specifically, we normalize these scores across various neighborhoods of closely contrasting questions and/or answers, adding a cross entropy loss term in addition to traditional maximum likelihood estimation. Our techniques require bundles of related question-answer pairs, which we either mine from within existing data or create using automated heuristics. We empirically demonstrate the effectiveness of training with instance bundles on two datasets—HotpotQA and ROPES—showing up to 9% absolute gains in accuracy.

Improving Query Graph Generation for Complex Question Answering over Knowledge Base

Kechen Qin et al.

Most of the existing Knowledge-based Question Answering (KBQA) methods first learn to map the given question to a query graph, and then convert the graph to an executable query to find the answer. The query graph is typically expanded progressively from the topic entity based on a sequence prediction model. In this paper, we propose a new solution to query graph generation that works in the opposite manner: we start with the entire knowledge base and gradually shrink it to the desired query graph. This approach improves both the efficiency and the accuracy of query graph generation, especially for complex multi-hop questions. Experimental results show that our method achieves state-of-the-art performance on ComplexWebQuestion (CWQ) dataset.

Evaluation Paradigms in Question Answering

Pedro Rodriguez and Jordan Boyd-Graber

Question answering (QA) primarily descends from two branches of research: (1) Alan Turing's investigation of machine intelligence at Manchester University and (2) Cyril Cleverdon's comparison of library card catalog indices at Cranfield University. This position paper names and distinguishes these paradigms. Despite substantial overlap, subtle but significant distinctions exert an outsize influence on research. While one evaluation

paradigm values creating more intelligent QA systems, the other paradigm values building QA systems that appeal to users. By better understanding the epistemic heritage of QA, researchers, academia, and industry can more effectively accelerate QA research.

Explaining Answers with Entailment Trees
Bhavana Dalvi et al.

Our goal, in the context of open-domain textual question-answering (QA), is to explain answers by showing the line of reasoning from what is known to the answer, rather than simply showing a fragment of textual evidence (a "rationale"). If this could be done, new opportunities for understanding and debugging the system's reasoning become possible. Our approach is to generate explanations in the form of entailment trees, namely a tree of multipremise entailment steps from facts that are known, through intermediate conclusions, to the hypothesis of interest (namely the question + answer). To train a model with this skill, we created ENTAILMENTBANK, the first dataset to contain multistep entailment trees. Given a hypothesis (question + answer), we define three increasingly difficult explanation tasks: generate a valid entailment tree given (a) all relevant sentences (b) all relevant and some irrelevant sentences, or (c) a corpus. We show that a strong language model can partially solve these tasks, in particular when the relevant sentences are included in the input (e.g., 35% of trees for (a) are perfect), and with indications of generalization to other domains. This work is significant as it provides a new type of dataset (multistep entailments) and baselines, offering a new avenue for the community to generate richer, more systematic explanations.

SituatedQA: Incorporating Extra-Linguistic Contexts into QA
Michael Zhang and Eunsol Choi

Answers to the same question may change depending on the extra-linguistic contexts (when and where the question was asked). To study this challenge, we introduce SituatedQA, an open-retrieval QA dataset where systems must produce the correct answer to a question given the temporal or geographical context. To construct SituatedQA, we first identify such questions in existing QA datasets. We find that a significant proportion of information seeking questions have context-dependent answers (e.g. roughly 16.5% of NQ-Open). For such context-dependent questions, we then crowdsource alternative contexts and their corresponding answers. Our study shows that existing models struggle with producing answers that are frequently updated or from uncommon locations. We further quantify how existing models, which are trained on data collected in the past, fail to generalize to answering questions asked in the present, even when provided with an updated evidence corpus (a roughly 15 point drop in accuracy). Our analysis suggests that open-retrieval QA benchmarks should incorporate extra-linguistic context to stay relevant globally and in the future. Our data, code, and datasheet are available at https://situatedqa.github.io/.

Surface Form Competition: Why the Highest Probability Answer Isn't Always Right
Ari Holtzman et al.

Large language models have shown promising results in zero-shot settings. For example, they can perform multiple choice tasks simply by conditioning on a question and selecting the answer with the highest probability. However, ranking by string probability can be problematic due to surface form competition—wherein different surface forms compete for probability mass, even if they represent the same underlying concept in a given context, e.g. "computer" and "PC." Since probability mass is finite, this lowers the probability of the correct answer, due to competition from other strings that are valid answers (but not one of the multiple choice options). We introduce Domain Conditional Pointwise Mutual Information, an alternative scoring function that directly compensates for surface form competition by simply reweighing each option according to its a priori likelihood within the context of a specific task. It achieves consistent gains in zero-shot performance over both calibrated and uncalibrated scoring functions on all GPT-2 and GPT-3 models on a variety of multiple choice datasets.

PAQ: 65 Million Probably-Asked Questions and What You Can Do With Them
Patrick Lewis et al.

Open-domain Question Answering models which directly leverage question-answer (QA) pairs, such as closed-book QA (CBQA) models and QA-pair retrievers, show promise in terms of speed and memory compared to conventional models which retrieve and read from text corpora. QA-pair retrievers also offer interpretable answers, a high degree of control, and are trivial to update at test time with new knowledge. However, these models fall short of the accuracy of retrieve-and-read systems, as substantially less knowledge is covered by the available QA-pairs relative to text corpora like Wikipedia. To facilitate improved QA-pair models, we introduce Probably Asked Questions (PAQ), a very large resource of 65M automatically-generated QA-pairs. We introduce a new QA-pair retriever, RePAQ, to complement PAQ. We find that PAQ pre-empts and caches test questions, enabling RePAQ to match the accuracy of recent retrieve-and-read models, whilst being significantly faster. Using PAQ, we train CBQA models which outperform comparable baselines by 5%, but trail RePAQ by over 15%, indicating the effectiveness of explicit retrieval. RePAQ can be configured for size (under 500MB) or speed (over 1K questions per second) whilst retaining high accuracy. Lastly, we demonstrate RePAQ's strength at selective QA, abstaining from answering when it is likely to be incorrect. This enables RePAQ to "back-off" to a more expensive state-of-the-art model, leading to a combined system which is both more accurate and 2x faster than the state-of-the-art model alone.

MS^2: Multi-Document Summarization of Medical Studies
Jay DeYoung et al.

To assess the effectiveness of any medical intervention, researchers must conduct a time-intensive and manual literature review. NLP systems can help to automate or assist in parts of this expensive process. In support of this goal, we release MS^2 (Multi-Document Summarization of Medical Studies), a dataset of over 470k documents and 20K summaries derived from the scientific literature. This dataset facilitates the development of systems that can assess and aggregate contradictory evidence across multiple studies, and is the first large-scale, publicly available multi-document summarization dataset in the biomedical domain. We experiment with a summarization system based on BART, with promising early results, though significant work remains to achieve higher summarization quality. We formulate our summarization inputs and targets in both free text and structured forms and modify a recently proposed metric to assess the quality of our system's generated summaries. Data and models are available at https://github.com/allenai/ms2.

CLIPScore: A Reference-free Evaluation Metric for Image Captioning

Jack Hessel et al.

Image captioning has conventionally relied on reference-based automatic evaluations, where machine captions are compared against captions written by humans. This is in contrast to the reference-free manner in which humans assess caption quality. In this paper, we report the surprising empirical finding that CLIP (Radford et al., 2021), a cross-modal model pretrained on 400M image+caption pairs from the web, can be used for robust automatic evaluation of image captioning without the need for references. Experiments spanning several corpora demonstrate that our new reference-free metric, CLIPScore, achieves the highest correlation with human judgements, outperforming existing reference-based metrics like CIDEr and SPICE. Information gain experiments demonstrate that CLIPScore, with its tight focus on image-text compatibility, is complementary to existing reference-based metrics that emphasize text-text similarities. Thus, we also present a reference-augmented version, RefCLIPScore, which achieves even higher correlation. Beyond literal description tasks, several case studies reveal domains where CLIPScore performs well (clip-art images, alt-text rating), but also where it is relatively weaker in comparison to reference-based metrics, e.g., news captions that require richer contextual knowledge.

Evaluating the Morphosyntactic Well-formedness of Generated Texts

Adithya Pratapa et al.

Text generation systems are ubiquitous in natural language processing applications. However, evaluation of these systems remains a challenge, especially in multilingual settings. In this paper, we propose L'AMBRE – a metric to evaluate the morphosyntactic well-formedness of text using its dependency parse and morphosyntactic rules of the language. We present a way to automatically extract various rules governing morphosyntax directly from dependency treebanks. To tackle the noisy outputs from text generation systems, we propose a simple methodology to train robust parsers. We show the effectiveness of our metric on the task of machine translation through a diachronic study of systems translating into morphologically-rich languages.

On the Influence of Masking Policies in Intermediate Pre-training

Qinyuan Ye et al.

Current NLP models are predominantly trained through a two-stage "pre-train then fine-tune" pipeline. Prior work has shown that inserting an intermediate pre-training stage, using heuristic masking policies for masked language modeling (MLM), can significantly improve final performance. However, it is still unclear (1) in what cases such intermediate pre-training is helpful, (2) whether hand-crafted heuristic objectives are optimal for a given task, and (3) whether a masking policy designed for one task is generalizable beyond that task. In this paper, we perform a large-scale empirical study to investigate the effect of various masking policies in intermediate pre-training with nine selected tasks across three categories. Crucially, we introduce methods to automate the discovery of optimal masking policies via direct supervision or meta-learning. We conclude that the success of intermediate pre-training is dependent on appropriate pre-train corpus, selection of output format (i.e., masked spans or full sentence), and clear understanding of the role that MLM plays for the downstream task. In addition, we find our learned masking policies outperform the heuristic of masking named entities on TriviaQA, and policies learned from one task can positively transfer to other tasks in certain cases, inviting future research in this direction.

On the Challenges of Evaluating Compositional Explanations in Multi-Hop Inference: Relevance, Completeness, and Expert Ratings

Peter Jansen et al.

Building compositional explanations requires models to combine two or more facts that, together, describe why the answer to a question is correct. Typically, these "multi-hop" explanations are evaluated relative to one (or a small number of) gold explanations. In this work, we show these evaluations substantially underestimate model performance, both in terms of the relevance of included facts, as well as the completeness of model-generated explanations, because models regularly discover and produce valid explanations that are different than gold explanations. To address this, we construct a large corpus of 126k domain-expert (science teacher) relevance ratings that augment a corpus of explanations to standardized science exam questions, discovering 80k additional relevant facts not rated as gold. We build three strong models based on different methodologies (generation, ranking, and schemas), and empirically show that while expert-augmented ratings provide better estimates of explanation quality, both original (gold) and expert-augmented automatic evaluations still substantially underestimate performance by up to 36% when compared with full manual expert judgements, with different models being disproportionately affected. This poses a significant methodological challenge to accurately evaluating explanations produced by compositional reasoning models.

ValNorm Quantifies Semantics to Reveal Consistent Valence Biases Across Languages and Over Centuries
Autumn Toney and Aylin Caliskan

Word embeddings learn implicit biases from linguistic regularities captured by word co-occurrence statistics. By extending methods that quantify human-like biases in word embeddings, we introduce ValNorm, a novel intrinsic evaluation task and method to quantify the valence dimension of affect in human-rated word sets from social psychology. We apply ValNorm on static word embeddings from seven languages (Chinese, English, German, Polish, Portuguese, Spanish, and Turkish) and from historical English text spanning 200 years. ValNorm achieves consistently high accuracy in quantifying the valence of non-discriminatory, non-social group word sets. Specifically, ValNorm achieves a Pearson correlation of r=0.88 for human judgment scores of valence for 399 words collected to establish pleasantness norms in English. In contrast, we measure gender stereotypes using the same set of word embeddings and find that social biases vary across languages. Our results indicate that valence associations of non-discriminatory, non-social group words represent widely-shared associations, in seven languages and over 200 years.

ESTER: A Machine Reading Comprehension Dataset for Reasoning about Event Semantic Relations
Rujun Han et al.

Understanding how events are semantically related to each other is the essence of reading comprehension. Recent event-centric reading comprehension datasets focus mostly on event arguments or temporal relations. While these tasks partially evaluate machines' ability of narrative understanding, human-like reading comprehension requires the capability to process event-based information beyond arguments and temporal reasoning. For example, to understand causality between events, we need to infer motivation or purpose; to establish event hierarchy, we need to understand the composition of events. To facilitate these tasks, we introduce **ESTER**, a comprehensive machine reading comprehension (MRC) dataset for Event Semantic Relation Reasoning. The dataset leverages natural language queries to reason about the five most common event semantic relations, provides more than 6K questions, and captures 10.1K event relation pairs. Experimental results show that the current SOTA systems achieve 22.1%, 63.3% and 83.5% for token-based exact-match (**EM**), **F1** and event-based **HIT1** scores, which are all significantly below human performances (36.0%, 79.6%, 100% respectively), highlighting our dataset as a challenging benchmark.

The Perils of Using Mechanical Turk to Evaluate Open-Ended Text Generation
Marzena Karpinska, Nader Akoury, and Mohit Iyyer

Recent text generation research has increasingly focused on open-ended domains such as story and poetry generation. Because models built for such tasks are difficult to evaluate automatically, most researchers in the space justify their modeling choices by collecting crowdsourced human judgments of text quality (e.g., Likert scores of coherence or grammaticality) from Amazon Mechanical Turk (AMT). In this paper, we first conduct a survey of 45 open-ended text generation papers and find that the vast majority of them fail to report crucial details about their AMT tasks, hindering reproducibility. We then run a series of story evaluation experiments with both AMT workers and English teachers and discover that even with strict qualification filters, AMT workers (unlike teachers) fail to distinguish between model-generated text and human-generated references. We show that AMT worker judgments improve when they are shown model-generated output alongside human-generated references, which enables the workers to better calibrate their ratings. Finally, interviews with the English teachers provide deeper insights into the challenges of the evaluation process, particularly when rating model-generated text.

RICA: Evaluating Robust Inference Capabilities Based on Commonsense Axioms
Pei Zhou et al.

Pre-trained language models (PTLMs) have achieved impressive performance on commonsense inference benchmarks, but their ability to employ commonsense to make robust inferences, which is crucial for effective communications with humans, is debated. In the pursuit of advancing fluid human-AI communication, we propose a new challenge, RICA: Robust Inference using Commonsense Axioms, that evaluates robust commonsense inference despite textual perturbations. To generate data for this challenge, we develop a systematic and scalable procedure using commonsense knowledge bases and probe PTLMs across two different evaluation settings. Extensive experiments on our generated probe sets with more than 10k statements show that PTLMs perform no better than random guessing on the zero-shot setting, are heavily impacted by statistical biases, and are not robust to perturbation attacks. We also find that fine-tuning on similar statements offer limited gains, as PTLMs still fail to generalize to unseen inferences. Our new large-scale benchmark exposes a significant gap between PTLMs and human-level language understanding and offers a new challenge for PTLMs to demonstrate commonsense.

Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus
Jesse Dodge et al.

Large language models have led to remarkable progress on many NLP tasks, and researchers are turning to ever-larger text corpora to train them. Some of the largest corpora available are made by scraping significant portions of the internet, and are frequently introduced with only minimal documentation. In this work we provide some of the first documentation for the Colossal Clean Crawled Corpus (C4; Raffel et al., 2020), a dataset created by applying a set of filters to a single snapshot of Common Crawl. We begin by investigating

where the data came from, and find a significant amount of text from unexpected sources like patents and US military websites. Then we explore the content of the text itself, and find machine-generated text (e.g., from machine translation systems) and evaluation examples from other benchmark NLP datasets. To understand the impact of the filters applied to create this dataset, we evaluate the text that was removed, and show that blocklist filtering disproportionately removes text from and about minority individuals. Finally, we conclude with some recommendations for how to created and document web-scale datasets from a scrape of the internet.

Evaluating the Evaluation Metrics for Style Transfer: A Case Study in Multilingual Formality Transfer
Eleftheria Briakou et al.

While the field of style transfer (ST) has been growing rapidly, it has been hampered by a lack of standardized practices for automatic evaluation. In this paper, we evaluate leading automatic metrics on the oft-researched task of formality style transfer. Unlike previous evaluations, which focus solely on English, we expand our focus to Brazilian-Portuguese, French, and Italian, making this work the first multilingual evaluation of metrics in ST. We outline best practices for automatic evaluation in (formality) style transfer and identify several models that correlate well with human judgments and are robust across languages. We hope that this work will help accelerate development in ST, where human evaluation is often challenging to collect.

Compression, Transduction, and Creation: A Unified Framework for Evaluating Natural Language Generation
Mingkai Deng et al.

Natural language generation (NLG) spans a broad range of tasks, each of which serves for specific objectives and desires different properties of generated text. The complexity makes automatic evaluation of NLG particularly challenging. Previous work has typically focused on a single task and developed individual evaluation metrics based on specific intuitions. In this paper, we propose a unifying perspective based on the nature of information change in NLG tasks, including compression (e.g., summarization), transduction (e.g., text rewriting), and creation (e.g., dialog). _Information alignment_ between input, context, and output text plays a common central role in characterizing the generation. With automatic alignment prediction models, we develop a family of interpretable metrics that are suitable for evaluating key aspects of different NLG tasks, often without need of gold reference data. Experiments show the uniformly designed metrics achieve stronger or comparable correlations with human judgement compared to state-of-the-art metrics in each of diverse tasks, including text summarization, style transfer, and knowledge-grounded dialog.

MS-Mentions: Consistently Annotating Entity Mentions in Materials Science Procedural Text
Tim O'Gorman et al.

Material science synthesis procedures are a promising domain for scientific NLP, as proper modeling of these recipes could provide insight into new ways of creating materials. However, a fundamental challenge in building information extraction models for material science synthesis procedures is getting accurate labels for the materials, operations, and other entities of those procedures. We present a new corpus of entity mention annotations over 595 Material Science synthesis procedural texts (157,488 tokens), which greatly expands the training data available for the Named Entity Recognition task. We outline a new label inventory designed to provide consistent annotations and a new annotation approach intended to maximize the consistency and annotation speed of domain experts. Inter-annotator agreement studies and baseline models trained upon the data suggest that the corpus provides high-quality annotations of these mention types. This corpus helps lay a foundation for future high-quality modeling of synthesis procedures.

The (Un)Suitability of Automatic Evaluation Metrics for Text Simplification
Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia

In order to simplify sentences, several rewriting operations can be performed, such as replacing complex words per simpler synonyms, deleting unnecessary information, and splitting long sentences. Despite this multi-operation nature, evaluation of automatic simplification systems relies on metrics that moderately correlate with human judgments on the simplicity achieved by executing specific operations (e.g., simplicity gain based on lexical replacements). In this article, we investigate how well existing metrics can assess sentence-level simplifications where multiple operations may have been applied and which, therefore, require more general simplicity judgments. For that, we first collect a new and more reliable data set for evaluating the correlation of metrics and human judgments of overall simplicity. Second, we conduct the first metaevaluation of automatic metrics in Text Simplification, using our new data set (and other existing data) to analyze the variation of the correlation between metrics' scores and human judgments across three dimensions: the perceived simplicity level, the system type, and the set of references used for computation. We show that these three aspects affect the correlations and, in particular, highlight the limitations of commonly used operation-specific metrics. Finally, based on our findings, we propose a set of recommendations for automatic evaluation of multi-operation simplifications, suggesting which metrics to compute and how to interpret their scores.

Memory-Based Semantic Parsing
Parag Jain and Mirella Lapata

We present a memory-based model for context-dependent semantic parsing. Previous approaches focus on enabling the decoder to copy or modify the parse from the previous utterance, assuming there is a dependency

between the current and previous parses. In this work, we propose to represent contextual information using an external memory. We learn a context memory controller that manages the memory by maintaining the cumulative meaning of sequential user utterances. We evaluate our approach on three semantic parsing benchmarks. Experimental results show that our model can better process context-dependent information and demonstrates improved performance without using task-specific decoders.

SimCSE: Simple Contrastive Learning of Sentence Embeddings
Tianyu Gao, Xingcheng Yao, and Danqi Chen

This paper presents SimCSE, a simple contrastive learning framework that greatly advances the state-of-the-art sentence embeddings. We first describe an unsupervised approach, which takes an input sentence and predicts itself in a contrastive objective, with only standard dropout used as noise. This simple method works surprisingly well, performing on par with previous supervised counterparts. We find that dropout acts as minimal data augmentation and removing it leads to a representation collapse. Then, we propose a supervised approach, which incorporates annotated pairs from natural language inference datasets into our contrastive learning framework, by using "entailment" pairs as positives and "contradiction" pairs as hard negatives. We evaluate SimCSE on standard semantic textual similarity (STS) tasks, and our unsupervised and supervised models using BERT base achieve an average of 76.3% and 81.6% Spearman's correlation respectively, a 4.2% and 2.2% improvement compared to previous best results. We also show—both theoretically and empirically—that contrastive learning objective regularizes pre-trained embeddings' anisotropic space to be more uniform, and it better aligns positive pairs when supervised signals are available.

Is Everything in Order? A Simple Way to Order Sentences
Somnath Basu Roy Chowdhury, Faeze Brahman, and Snigdha Chaturvedi

The task of organizing a shuffled set of sentences into a coherent text has been used to evaluate a machine's understanding of causal and temporal relations. We formulate the sentence ordering task as a conditional text-to-marker generation problem. We present Reorder-BART (Re-BART) that leverages a pre-trained Transformer-based model to identify a coherent order for a given set of shuffled sentences. The model takes a set of shuffled sentences with sentence-specific markers as input and generates a sequence of position markers of the sentences in the ordered text. Re-BART achieves the state-of-the-art performance across 7 datasets in Perfect Match Ratio (PMR) and Kendall's tau. We perform evaluations in a zero-shot setting, showcasing that our model is able to generalize well across other datasets. We additionally perform several experiments to understand the functioning and limitations of our framework.

When is Wall a Pared and when a Muro?: Extracting Rules Governing Lexical Selection
Aditi Chaudhary et al.

Learning fine-grained distinctions between vocabulary items is a key challenge in learning a new language. For example, the noun "wall" has different lexical manifestations in Spanish – "pared" refers to an indoor wall while "muro" refers to an outside wall. However, this variety of lexical distinction may not be obvious to non-native learners unless the distinction is explained in such a way. In this work, we present a method for automatically identifying fine-grained lexical distinctions, and extracting rules explaining these distinctions in a human- and machine-readable format. We confirm the quality of these extracted rules in a language learning setup for two languages, Spanish and Greek, where we use the rules to teach non-native speakers when to translate a given ambiguous word into its different possible translations.

Controllable Semantic Parsing via Retrieval Augmentation
Panupong Pasupat, Yuan Zhang, and Kelvin Guu

In practical applications of semantic parsing, we often want to rapidly change the behavior of the parser, such as enabling it to handle queries in a new domain, or changing its predictions on certain targeted queries. While we can introduce new training examples exhibiting the target behavior, a mechanism for enacting such behavior changes without expensive model re-training would be preferable. To this end, we propose ControllAble Semantic Parser via Exemplar Retrieval (CASPER). Given an input query, the parser retrieves related exemplars from a retrieval index, augments them to the query, and then applies a generative seq2seq model to produce an output parse. The exemplars act as a control mechanism over the generic generative model: by manipulating the retrieval index or how the augmented query is constructed, we can manipulate the behavior of the parser. On the MTOP dataset, in addition to achieving state-of-the-art on the standard setup, we show that CASPER can parse queries in a new domain, adapt the prediction toward the specified patterns, or adapt to new semantic schemas without having to further re-train the model.

Constrained Language Models Yield Few-Shot Semantic Parsers
Richard Shin et al.

We explore the use of large pretrained language models as few-shot semantic parsers. The goal in semantic parsing is to generate a structured meaning representation given a natural language input. However, language models are trained to generate natural language. To bridge the gap, we use language models to paraphrase inputs into a controlled sublanguage resembling English that can be automatically mapped to a target meaning representation. Our results demonstrate that with only a small amount of data and very little code to convert into English-like representations, our blueprint for rapidly bootstrapping semantic parsers leads to surprisingly effective performance on multiple community tasks, greatly exceeding baseline methods also trained on the same limited data.

Phrase-BERT: Improved Phrase Embeddings from BERT with an Application to Corpus Exploration
Shufan Wang, Laure Thompson, and Mohit Iyyer

Phrase representations derived from BERT often do not exhibit complex phrasal compositionality, as the model relies instead on lexical similarity to determine semantic relatedness. In this paper, we propose a contrastive fine-tuning objective that enables BERT to produce more powerful phrase embeddings. Our approach (Phrase-BERT) relies on a dataset of diverse phrasal paraphrases, which is automatically generated using a paraphrase generation model, as well as a large-scale dataset of phrases in context mined from the Books3 corpus. Phrase-BERT outperforms baselines across a variety of phrase-level similarity tasks, while also demonstrating increased lexical diversity between nearest neighbors in the vector space. Finally, as a case study, we show that Phrase-BERT embeddings can be easily integrated with a simple autoencoder to build a phrase-based neural topic model that interprets topics as mixtures of words and phrases by performing a nearest neighbor search in the embedding space. Crowdsourced evaluations demonstrate that this phrase-based topic model produces more coherent and meaningful topics than baseline word and phrase-level topic models, further validating the utility of Phrase-BERT.

ExplaGraphs: An Explanation Graph Generation Task for Structured Commonsense Reasoning
Swarnadeep Saha et al.

Recent commonsense-reasoning tasks are typically discriminative in nature, where a model answers a multiple-choice question for a certain context. Discriminative tasks are limiting because they fail to adequately evaluate the model's ability to reason and explain predictions with underlying commonsense knowledge. They also allow such models to use reasoning shortcuts and not be "right for the right reasons". In this work, we present ExplaGraphs, a new generative and structured commonsense-reasoning task (and an associated dataset) of explanation graph generation for stance prediction. Specifically, given a belief and an argument, a model has to predict if the argument supports or counters the belief and also generate a commonsense-augmented graph that serves as non-trivial, complete, and unambiguous explanation for the predicted stance. We collect explanation graphs through a novel Create-Verify-And-Refine graph collection framework that improves the graph quality (up to 90%) via multiple rounds of verification and refinement. A significant 79% of our graphs contain external commonsense nodes with diverse structures and reasoning depths. Next, we propose a multi-level evaluation framework, consisting of automatic metrics and human evaluation, that check for the structural and semantic correctness of the generated graphs and their degree of match with ground-truth graphs. Finally, we present several structured, commonsense-augmented, and text generation models as strong starting points for this explanation graph generation task, and observe that there is a large gap with human performance, thereby encouraging future work for this new challenging task.

Connect-the-Dots: Bridging Semantics between Words and Definitions via Aligning Word Sense Inventories
Wenlin Yao et al.

Word Sense Disambiguation (WSD) aims to automatically identify the exact meaning of one word according to its context. Existing supervised models struggle to make correct predictions on rare word senses due to limited training data and can only select the best definition sentence from one predefined word sense inventory (e.g., WordNet). To address the data sparsity problem and generalize the model to be independent of one predefined inventory, we propose a gloss alignment algorithm that can align definition sentences (glosses) with the same meaning from different sense inventories to collect rich lexical knowledge. We then train a model to identify semantic equivalence between a target word in context and one of its glosses using these aligned inventories, which exhibits strong transfer capability to many WSD tasks. Experiments on benchmark datasets show that the proposed method improves predictions on both frequent and rare word senses, outperforming prior work by 1.2% on the All-Words WSD Task and 4.3% on the Low-Shot WSD Task. Evaluation on WiC Task also indicates that our method can better capture word meanings in context.

Numeracy enhances the Literacy of Language Models
Avijit Thawani, Jay Pujara, and Filip Ilievski

Specialized number representations in NLP have shown improvements on numerical reasoning tasks like arithmetic word problems and masked number prediction. But humans also use numeracy to make better sense of world concepts, e.g., you can seat 5 people in your 'room' but not 500. Does a better grasp of numbers improve a model's understanding of other concepts and words? This paper studies the effect of using six different number encoders on the task of masked word prediction (MWP), as a proxy for evaluating literacy. To support this investigation, we develop Wiki-Convert, a 900,000 sentence dataset annotated with numbers and units, to avoid conflating nominal and ordinal number occurrences. We find a significant improvement in MWP for sentences containing numbers, that exponent embeddings are the best number encoders, yielding over 2 points jump in prediction accuracy over a BERT baseline, and that these enhanced literacy skills also generalize to contexts without annotated numbers. We release all code at https://git.io/JuZXn.

Powering Comparative Classification with Sentiment Analysis via Domain Adaptive Knowledge Transfer
Zeyu Li et al.

We study Comparative Preference Classification (CPC) which aims at predicting whether a preference com-

parison exists between two entities in a given sentence and, if so, which entity is preferred over the other. High-quality CPC models can significantly benefit applications such as comparative question answering and review-based recommendation. Among the existing approaches, non-deep learning methods suffer from inferior performances. The state-of-the-art graph neural network-based ED-GAT (Ma et al., 2020) only considers syntactic information while ignoring the critical semantic relations and the sentiments to the compared entities. We propose Sentiment Analysis Enhanced COmparative Network (SAECON) which improves CPC accuracy with a sentiment analyzer that learns sentiments to individual entities via domain adaptive knowledge transfer. Experiments on the CompSent-19 (Panchenko et al., 2019) dataset present a significant improvement on the F1 scores over the best existing CPC approaches.

SYSML: StYlometry with Structure and Multitask Learning: Implications for Darknet Forum Migrant Analysis
Pranav Maneriker, Yuntian He, and Srinivasan Parthasarathy

Darknet market forums are frequently used to exchange illegal goods and services between parties who use encryption to conceal their identities. The Tor network is used to host these markets, which guarantees additional anonymization from IP and location tracking, making it challenging to link across malicious users using multiple accounts (sybils). Additionally, users migrate to new forums when one is closed further increasing the difficulty of linking users across multiple forums. We develop a novel stylometry-based multitask learning approach for natural language and model interactions using graph embeddings to construct low-dimensional representations of short episodes of user activity for authorship attribution. We provide a comprehensive evaluation of our methods across four different darknet forums demonstrating its efficacy over the state-of-the-art, with a lift of up to 2.5X on Mean Retrieval Rank and 2X on Recall10.

CHoRaL: Collecting Humor Reaction Labels from Millions of Social Media Users
Zixiaofan Yang, Shayan Hooshmand, and Julia Hirschberg

Humor detection has gained attention in recent years due to the desire to understand user-generated content with figurative language. However, substantial individual and cultural differences in humor perception make it very difficult to collect a large-scale humor dataset with reliable humor labels. We propose CHoRaL, a framework to generate perceived humor labels on Facebook posts, using the naturally available user reactions to these posts with no manual annotation needed. CHoRaL provides both binary labels and continuous scores of humor and non-humor. We present the largest dataset to date with labeled humor on 785K posts related to COVID-19. Additionally, we analyze the expression of COVID-related humor in social media by extracting lexico-semantic and affective features from the posts, and build humor detection models with performance similar to humans. CHoRaL enables the development of large-scale humor detection models on any topic and opens a new path to the study of humor on social media.

CLASSIC: Continual and Contrastive Learning of Aspect Sentiment Classification Tasks
Zixuan Ke et al.

This paper studies continual learning (CL) of a sequence of aspect sentiment classification (ASC) tasks in a particular CL setting called domain incremental learning (DIL). Each task is from a different domain or product. The DIL setting is particularly suited to ASC because in testing the system needs not know the task/domain to which the test data belongs. To our knowledge, this setting has not been studied before for ASC. This paper proposes a novel model called CLASSIC. The key novelty is a contrastive continual learning method that enables both knowledge transfer across tasks and knowledge distillation from old tasks to the new task, which eliminates the need for task ids in testing. Experimental results show the high effectiveness of CLASSIC.

Visual Goal-Step Inference using wikiHow
Yue Yang et al.

Understanding what sequence of steps are needed to complete a goal can help artificial intelligence systems reason about human activities. Past work in NLP has examined the task of goal-step inference for text. We introduce the visual analogue. We propose the Visual Goal-Step Inference (VGSI) task, where a model is given a textual goal and must choose which of four images represents a plausible step towards that goal. With a new dataset harvested from wikiHow consisting of 772,277 images representing human actions, we show that our task is challenging for state-of-the-art multimodal models. Moreover, the multimodal representation learned from our data can be effectively transferred to other datasets like HowTo100m, increasing the VGSI accuracy by 15 - 20%. Our task will facilitate multimodal reasoning about procedural events.

Visual News: Benchmark and Challenges in News Image Captioning
Fuxiao Liu et al.

We propose Visual News Captioner, an entity-aware model for the task of news image captioning. We also introduce Visual News, a large-scale benchmark consisting of more than one million news images along with associated news articles, image captions, author information, and other metadata. Unlike the standard image captioning task, news images depict situations where people, locations, and events are of paramount importance. Our proposed method can effectively combine visual and textual features to generate captions with richer information such as events and entities. More specifically, built upon the Transformer architecture, our model is further equipped with novel multi-modal feature fusion techniques and attention mechanisms, which are designed to generate named entities more accurately. Our method utilizes much fewer parameters while

achieving slightly better prediction results than competing methods. Our larger and more diverse Visual News dataset further highlights the remaining challenges in captioning news images.

WhyAct: Identifying Action Reasons in Lifestyle Vlogs
Oana Ignat et al.

We aim to automatically identify human action reasons in online videos. We focus on the widespread genre of lifestyle vlogs, in which people perform actions while verbally describing them. We introduce and make publicly available the WhyAct dataset, consisting of 1,077 visual actions manually annotated with their reasons. We describe a multimodal model that leverages visual and textual information to automatically infer the reasons corresponding to an action presented in the video.

CrossVQA: Scalably Generating Benchmarks for Systematically Testing VQA Generalization
Arjun Akula et al.

One challenge in evaluating visual question answering (VQA) models in the cross-dataset adaptation setting is that the distribution shifts are multi-modal, making it difficult to identify if it is the shifts in visual or language features that play a key role. In this paper, we propose a semi-automatic framework for generating disentangled shifts by introducing a controllable visual question-answer generation (VQAG) module that is capable of generating highly-relevant and diverse question-answer pairs with the desired dataset style. We use it to create CrossVQA, a collection of test splits for assessing VQA generalization based on the VQA2, VizWiz, and Open~Images datasets. We provide an analysis of our generated datasets and demonstrate its utility by using them to evaluate several state-of-the-art VQA systems. One important finding is that the visual shifts in cross-dataset VQA matter more than the language shifts. More broadly, we present a scalable framework for systematically evaluating the machine with little human intervention.

Systematic Generalization on gSCAN: What is Nearly Solved and What is Next?
Linlu Qiu et al.

We analyze the *grounded* SCAN (gSCAN) benchmark, which was recently proposed to study systematic generalization for grounded language understanding. First, we study which aspects of the original benchmark can be solved by commonly used methods in multi-modal research. We find that a general-purpose Transformer-based model with cross-modal attention achieves strong performance on a majority of the gSCAN splits, surprisingly outperforming more specialized approaches from prior work. Furthermore, our analysis suggests that many of the remaining errors reveal the same fundamental challenge in systematic generalization of linguistic constructs regardless of visual context. Second, inspired by this finding, we propose challenging new tasks for gSCAN by generating data to incorporate relations between objects in the visual environment. Finally, we find that current models are surprisingly data inefficient given the narrow scope of commands in gSCAN, suggesting another challenge for future work.

Integrating Visuospatial, Linguistic, and Commonsense Structure into Story Visualization
Adyasha Maharana and Mohit Bansal

While much research has been done in text-to-image synthesis, little work has been done to explore the usage of linguistic structure of the input text. Such information is even more important for story visualization since its inputs have an explicit narrative structure that needs to be translated into an image sequence (or visual story). Prior work in this domain has shown that there is ample room for improvement in the generated image sequence in terms of visual quality, consistency and relevance. In this paper, we first explore the use of constituency parse trees using a Transformer-based recurrent architecture for encoding structured input. Second, we augment the structured input with commonsense information and study the impact of this external knowledge on the generation of visual story. Third, we also incorporate visual structure via bounding boxes and dense captioning to provide feedback about the characters/objects in generated images within a dual learning setup. We show that off-the-shelf dense-captioning models trained on Visual Genome can improve the spatial structure of images from a different target domain without needing fine-tuning. We train the model end-to-end using intra-story contrastive loss (between words and image sub-regions) and show significant improvements in visual quality. Finally, we provide an analysis of the linguistic and visuo-spatial information.

VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding
Hu Xu et al.

We present VideoCLIP, a contrastive approach to pre-train a unified model for zero-shot video and text understanding, without using any labels on downstream tasks. VideoCLIP trains a transformer for video and text by contrasting temporally overlapping positive video-text pairs with hard negatives from nearest neighbor retrieval. Our experiments on a diverse series of downstream tasks, including sequence-level text-video retrieval, VideoQA, token-level action localization, and action segmentation reveal state-of-the-art performance, surpassing prior work, and in some cases even outperforming supervised approaches. Code is made available at https://github.com/pytorch/fairseq/examples/MMPT.

NewsCLIPpings: Automatic Generation of Out-of-Context Multimodal Media
Grace Luo, Trevor Darrell, and Anna Rohrbach

Online misinformation is a prevalent societal issue, with adversaries relying on tools ranging from cheap fakes to sophisticated deep fakes. We are motivated by the threat scenario where an image is used out of context to support a certain narrative. While some prior datasets for detecting image-text inconsistency generate samples via text manipulation, we propose a dataset where both image and text are unmanipulated but mismatched. We introduce several strategies for automatically retrieving convincing images for a given caption, capturing cases with inconsistent entities or semantic context. Our large-scale automatically generated the NewsCLIPpings Dataset: (1) demonstrates that machine-driven image repurposing is now a realistic threat, and (2) provides samples that represent challenging instances of mismatch between text and image in news that are able to mislead humans. We benchmark several state-of-the-art multimodal models on our dataset and analyze their performance across different pretraining domains and visual backbones.

Multimodal Pretraining Unmasked: A Meta-Analysis and a Unified Framework of Vision-and-Language BERTs
Emanuele Bugliarello et al.

Large-scale pretraining and task-specific fine-tuning is now the standard methodology for many tasks in computer vision and natural language processing. Recently, a multitude of methods have been proposed for pretraining vision and language BERTs to tackle challenges at the intersection of these two key areas of AI. These models can be categorised into either single-stream or dual-stream encoders. We study the differences between these two categories, and show how they can be unified under a single theoretical framework. We then conduct controlled experiments to discern the empirical differences between five VL BERTs. Our experiments show that training data and hyperparameters are responsible for most of the differences between the reported results, but they also reveal that the embedding layer plays a crucial role in these massive models.

Multiplex Graph Neural Network for Extractive Text Summarization
Baoyu Jing et al.

Extractive text summarization aims at extracting the most representative sentences from a given document as its summary. To extract a good summary from a long text document, sentence embedding plays an important role. Recent studies have leveraged graph neural networks to capture the inter-sentential relationship (e.g., the discourse graph) within the documents to learn contextual sentence embedding. However, those approaches neither consider multiple types of inter-sentential relationships (e.g., semantic similarity and natural connection relationships), nor model intra-sentential relationships (e.g, semantic similarity and syntactic relationship among words). To address these problems, we propose a novel Multiplex Graph Convolutional Network (Multi-GCN) to jointly model different types of relationships among sentences and words. Based on Multi-GCN, we propose a Multiplex Graph Summarization (Multi-GraS) model for extractive text summarization. Finally, we evaluate the proposed models on the CNN/DailyMail benchmark dataset to demonstrate effectiveness of our method.

HETFORMER: Heterogeneous Transformer with Sparse Attention for Long-Text Extractive Summarization
Ye Liu et al.

To capture the semantic graph structure from raw text, most existing summarization approaches are built on GNNs with a pre-trained model. However, these methods suffer from cumbersome procedures and inefficient computations for long-text documents. To mitigate these issues, this paper proposes HetFormer, a Transformer-based pre-trained model with multi-granularity sparse attentions for long-text extractive summarization. Specifically, we model different types of semantic nodes in raw text as a potential heterogeneous graph and directly learn heterogeneous relationships (edges) among nodes by Transformer. Extensive experiments on both single- and multi-document summarization tasks show that HetFormer achieves state-of-the-art performance in Rouge F1 while using less memory and fewer parameters.

Simple Conversational Data Augmentation for Semi-supervised Abstractive Dialogue Summarization
Jiaao Chen and Diyi Yang

Abstractive conversation summarization has received growing attention while most current state-of-the-art summarization models heavily rely on human-annotated summaries. To reduce the dependence on labeled summaries, in this work, we present a simple yet effective set of Conversational Data Augmentation (CODA) methods for semi-supervised abstractive conversation summarization, such as random swapping/deletion to perturb the discourse relations inside conversations, dialogue-acts-guided insertion to interrupt the development of conversations, and conditional-generation-based substitution to substitute utterances with their paraphrases generated based on the conversation context. To further utilize unlabeled conversations, we combine CODA with two-stage noisy self-training where we first pre-train the summarization model on unlabeled conversations with pseudo summaries and then fine-tune it on labeled conversations. Experiments conducted on the recent conversation summarization datasets demonstrate the effectiveness of our methods over several state-of-the-art data augmentation baselines.

Finding a Balanced Degree of Automation for Summary Evaluation
Shiyue Zhang and Mohit Bansal

Human evaluation for summarization tasks is reliable but brings in issues of reproducibility and high costs. Automatic metrics are cheap and reproducible but sometimes poorly correlated with human judgment. In this

work, we propose flexible semiautomatic to automatic summary evaluation metrics, following the Pyramid human evaluation method. Semi-automatic Lite2Pyramid retains the reusable human-labeled Summary Content Units (SCUs) for reference(s) but replaces the manual work of judging SCUs' presence in system summaries with a natural language inference (NLI) model. Fully automatic Lite3Pyramid further substitutes SCUs with automatically extracted Semantic Triplet Units (STUs) via a semantic role labeling (SRL) model. Finally, we propose in-between metrics, Lite2.xPyramid, where we use a simple regressor to predict how well the STUs can simulate SCUs and retain SCUs that are more difficult to simulate, which provides a smooth transition and balance between automation and manual evaluation. Comparing to 15 existing metrics, we evaluate human-metric correlations on 3 existing meta-evaluation datasets and our newly collected PyrXSum (with 100/10 XSum examples/systems). It shows that Lite2Pyramid consistently has the best summary-level correlations; Lite3Pyramid works better than or comparable to other automatic metrics; Lite2.xPyramid trades off small correlation drops for larger manual effort reduction, which can reduce costs for future data collection.

Decision-Focused Summarization

Chao-Chun Hsu and Chenhao Tan

Relevance in summarization is typically de- fined based on textual information alone, without incorporating insights about a particular decision. As a result, to support risk analysis of pancreatic cancer, summaries of medical notes may include irrelevant information such as a knee injury. We propose a novel problem, decision-focused summarization, where the goal is to summarize relevant information for a decision. We leverage a predictive model that makes the decision based on the full text to provide valuable insights on how a decision can be inferred from text. To build a summary, we then select representative sentences that lead to similar model decisions as using the full text while accounting for textual non-redundancy. To evaluate our method (DecSum), we build a testbed where the task is to summarize the first ten reviews of a restaurant in support of predicting its future rating on Yelp. DecSum substantially outperforms text-only summarization methods and model-based explanation methods in decision faithfulness and representativeness. We further demonstrate that DecSum is the only method that enables humans to outperform random chance in predicting which restaurant will be better rated in the future.

AUTOSUMM: Automatic Model Creation for Text Summarization

Sharmila Reddy Nangi et al.

Recent efforts to develop deep learning models for text generation tasks such as extractive and abstractive summarization have resulted in state-of-the-art performances on various datasets. However, obtaining the best model configuration for a given dataset requires an extensive knowledge of deep learning specifics like model architecture, tuning parameters etc., and is often extremely challenging for a non-expert. In this paper, we propose methods to automatically create deep learning models for the tasks of extractive and abstractive text summarization. Based on the recent advances in Automated Machine Learning and the success of large language models such as BERT and GPT-2 in encoding knowledge, we use a combination of Neural Architecture Search (NAS) and Knowledge Distillation (KD) techniques to perform model search and compression using the vast knowledge provided by these language models to develop smaller, customized models for any given dataset. We present extensive empirical results to illustrate the effectiveness of our model creation methods in terms of inference time and model size, while achieving near state-of-the-art performances in terms of accuracy across a range of datasets.

Genre as Weak Supervision for Cross-lingual Dependency Parsing

Max Müller-Eberstein, Rob van der Goot, and Barbara Plank

Recent work has shown that monolingual masked language models learn to represent data-driven notions of language variation which can be used for domain-targeted training data selection. Dataset genre labels are already frequently available, yet remain largely unexplored in cross-lingual setups. We harness this genre metadata as a weak supervision signal for targeted data selection in zero-shot dependency parsing. Specifically, we project treebank-level genre information to the finer-grained sentence level, with the goal to amplify information implicitly stored in unsupervised contextualized representations. We demonstrate that genre is recoverable from multilingual contextual embeddings and that it provides an effective signal for training data selection in cross-lingual, zero-shot scenarios. For 12 low-resource language treebanks, six of which are test-only, our genre-specific methods significantly outperform competitive baselines as well as recent embedding-based methods for data selection. Moreover, genre-based data selection provides new state-of-the-art results for three of these target languages.

Improved Latent Tree Induction with Distant Supervision via Span Constraints

Zhiyang Xu et al.

For over thirty years, researchers have developed and analyzed methods for latent tree induction as an approach for unsupervised syntactic parsing. Nonetheless, modern systems still do not perform well enough compared to their supervised counterparts to have any practical use as structural annotation of text. In this work, we present a technique that uses distant supervision in the form of span constraints (i.e. phrase bracketing) to improve performance in unsupervised constituency parsing. Using a relatively small number of span constraints we can substantially improve the output from DIORA, an already competitive unsupervised parsing system. Compared with full parse tree annotation, span constraints can be acquired with minimal effort, such as with a lexicon derived from Wikipedia, to find exact text matches. Our experiments show span constraints based on entities improves constituency parsing on English WSJ Penn Treebank by more than 5 F1. Furthermore, our method extends to any domain where span constraints are easily attainable, and as a case study we demonstrate its

effectiveness by parsing biomedical text from the CRAFT dataset.

Virtual Demo Session

Automatic Construction of Enterprise Knowledge Base
Junyi Chai et al.

In this paper, we present an automatic knowledge base construction system from large scale enterprise documents with minimal efforts of human intervention. In the design and deployment of such a knowledge mining system for enterprise, we faced several challenges including data distributional shift, performance evaluation, compliance requirements and other practical issues. We leveraged state-of-the-art deep learning models to extract information (named entities and definitions) at per document level, then further applied classical machine learning techniques to process global statistical information to improve the knowledge base. Experimental results are reported on actual enterprise documents. This system is currently serving as part of a Microsoft 365 service.

LightTag: Text Annotation Platform
Tal Perry

Text annotation tools assume that their user goal is to create a labeled corpus. However, users view annotation as a necessary evil on the way to deliver business value through NLP. Thus an annotation tool should optimize for the throughput of the global NLP process, not only the productivity of individual annotators. LightTag is a text annotation tool designed and built on that principle. This paper shares our design rationale, data modeling choices, and user interface decisions then illustrates how those choices serve the full NLP lifecycle.

ET: A Workstation for Querying, Editing and Evaluating Annotated Corpora
Elvis de Souza and Cláudia Freitas

In this paper we explore the functionalities of ET, a suite designed to support linguistic research and natural language processing tasks using corpora annotated in the CoNLL-U format. These goals are achieved by two integrated environments Interrogatorio, an environment for querying and editing annotated corpora, and Julgamento, an environment for assessing their quality. ET is open-source, built on different Python Web technologies and has Web demonstrations available on-line. ET has been intensively used in our research group for over two years, being the chosen framework for several linguistic and NLP-related studies conducted by its researchers.

COMBO: State-of-the-Art Morphosyntactic Analysis
Mateusz Klimaszewski and Alina Wróblewska

We introduce COMBO, a fully neural NLP system for accurate part-of-speech tagging, morphological analysis, lemmatisation, and (enhanced) dependency parsing. It predicts categorical morphosyntactic features whilst also exposes their vector representations, extracted from hidden layers. COMBO is an easy to install Python package with automatically downloadable pre-trained models for over 40 languages. It maintains a balance between efficiency and quality. As it is an end-to-end system and its modules are jointly trained, its training is competitively fast. As its models are optimised for accuracy, they achieve often better prediction quality than SOTA. The COMBO library is available at: https://gitlab.clarin-pl.eu/syntactic-tools/combo.

ExcavatorCovid: Extracting Events and Relations from Text Corpora for Temporal and Causal Analysis for COVID-19
Bonan Min et al.

Timely responses from policy makers to mitigate the impact of the COVID-19 pandemic rely on a comprehensive grasp of events, their causes, and their impacts. These events are reported at such a speed and scale as to be overwhelming. In this paper, we present ExcavatorCovid, a machine reading system that ingests open-source text documents (e.g., news and scientific publications), extracts COVID-19 related events and relations between them, and builds a Temporal and Causal Analysis Graph (TCAG). Excavator will help government agencies alleviate the information overload, understand likely downstream effects of political and economic decisions and events related to the pandemic, and respond in a timely manner to mitigate the impact of COVID-19. We expect the utility of Excavator to outlive the COVID-19 pandemic: analysts and decision makers will be empowered by Excavator to better understand and solve complex problems in the future. A demonstration video is available at https://vimeo.com/528619007.

RepGraph: Visualising and Analysing Meaning Representation Graphs
Jaron Cohen et al.

We present RepGraph, an open source visualisation and analysis tool for meaning representation graphs. Graph-based meaning representations provide rich semantic annotations, but visualising them clearly is more challenging than for fully lexicalized representations. Our application provides a seamless, unifying interface with which to visualise, manipulate and analyse semantically parsed graph data represented in a JSON-based serialisation format. RepGraph visualises graphs in multiple formats, with an emphasis on showing the relation between nodes and their corresponding token spans, whilst keeping the representation compact. Additionally, the web-based tool provides NLP researchers with a clear, visually intuitive way of interacting with these graphs, and includes a number of graph analysis features. The tool currently supports the DMRS, EDS, PTG, UCCA, and AMR semantic frameworks. A live demo is available at https://repgraph.vercel.app/.

Thermostat: A Large Collection of NLP Model Explanations and Analysis Tools
Nils Feldhus, Robert Schwarzenberg, and Sebastian Möller

In the language domain, as in other domains, neural explainability takes an ever more important role, with feature attribution methods on the forefront. Many such methods require considerable computational resources and expert knowledge about implementation details and parameter choices. To facilitate research, we present Thermostat which consists of a large collection of model explanations and accompanying analysis tools. Thermostat allows easy access to over 200k explanations for the decisions of prominent state-of-the-art models spanning across different NLP tasks, generated with multiple explainers. The dataset took over 10k GPU hours (> one year) to compile; compute time that the community now saves. The accompanying software tools allow to analyse explanations instance-wise but also accumulatively on corpus level. Users can investigate and compare models, datasets and explainers without the need to orchestrate implementation details. Thermostat is fully open source, democratizes explainability research in the language domain, circumvents redundant computations and increases comparability and replicability.

LMdiff: A Visual Diff Tool to Compare Language Models
Hendrik Strobelt et al.

While different language models are ubiquitous in NLP, it is hard to contrast their outputs and identify which contexts one can handle better than the other. To address this question, we introduce LMdiff, a tool that visually compares probability distributions of two models that differ, e.g., through finetuning, distillation, or simply training with different parameter sizes. LMdiff allows the generation of hypotheses about model behavior by investigating text instances token by token and further assists in choosing these interesting text instances by identifying the most interesting phrases from large corpora. We showcase the applicability of LMdiff for hypothesis generation across multiple case studies. A demo is available at http://lmdiff.net .

Semantic Context Path Labeling for Semantic Exploration of User Reviews
Salah Aït-Mokhtar et al.

In this paper we present a prototype demonstrator showcasing a novel method to perform semantic exploration of user reviews. The system enables effective navigation in a rich contextual semantic schema with a large number of structured classes indicating relevant information. In order to identify instances of the structured classes in the reviews, we defined a new Information Extraction task called Semantic Context Path (SCP) labeling, which simultaneously assigns types and semantic roles to entity mentions. Reviews can rapidly be explored based on the fine-grained and structured semantic classes. As a proof-of-concept, we have implemented this system for reviews on Points-of-Interest, in English and Korean.

Beyond Accuracy: A Consolidated Tool for Visual Question Answering Benchmarking
Dirk Väth, Pascal Tilli, and Ngoc Thang Vu

On the way towards general Visual Question Answering (VQA) systems that are able to answer arbitrary questions, the need arises for evaluation beyond single-metric leaderboards for specific datasets. To this end, we propose a browser-based benchmarking tool for researchers and challenge organizers, with an API for easy integration of new models and datasets to keep up with the fast-changing landscape of VQA. Our tool helps test generalization capabilities of models across multiple datasets, evaluating not just accuracy, but also performance in more realistic real-world scenarios such as robustness to input noise. Additionally, we include metrics that measure biases and uncertainty, to further explain model behavior. Interactive filtering facilitates discovery of problematic behavior, down to the data sample level. As proof of concept, we perform a case study on four models. We find that state-of-the-art VQA models are optimized for specific tasks or datasets, but fail to generalize even to other in-domain test sets, for example they can not recognize text in images. Our metrics allow us to quantify which image and question embeddings provide most robustness to a model. All code s publicly available.

SPRING Goes Online: End-to-End AMR Parsing and Generation
Rexhina Blloshmi et al.

In this paper we present SPRING Online Services, a Web interface and RESTful APIs for our state-of-the-art AMR parsing and generation system, SPRING (Symmetric PaRsIng aNd Generation). The Web interface has been developed to be easily used by the Natural Language Processing community, as well as by the general public. It provides, among other things, a highly interactive visualization platform and a feedback mechanism to obtain user suggestions for further improvements of the system's output. Moreover, our RESTful APIs enable easy integration of SPRING in downstream applications where AMR structures are needed. Finally, we make SPRING Online Services freely available at http://nlp.uniroma1.it/spring and, in addition, we release extra model checkpoints to be used with the original SPRING Python code.

Press Freedom Monitor: Detection of Reported Press and Media Freedom Violations in Twitter and News Articles
Tariq Yousef et al.

Freedom of the press and media is of vital importance for democratically organised states and open societies. We introduce the Press Freedom Monitor, a tool that aims to detect reported press and media freedom violations in news articles and tweets. It is used by press and media freedom organisations to support their daily monitoring and to trigger rapid response actions. The Press Freedom Monitor enables the monitoring experts to get a fast overview over recently reported incidents and it has shown an impressive performance in this

regard. This paper presents our work on the tool, starting with the training phase, which comprises defining the topic-related keywords to be used for querying APIs for news and Twitter content and evaluating different machine learning models based on a training dataset specifically created for our use case. Then, we describe the components of the production pipeline, including data gathering, duplicates removal, country mapping, case mapping and the user interface. We also conducted a usability study to evaluate the effectiveness of the user interface, and describe improvement plans for future work.

TranslateLocally: Blazing-fast translation running on the local CPU
Nikolay Bogoychev, Jelmer Van der Linde, and Kenneth Heafield

Every day, millions of people sacrifice their privacy and browsing habits in exchange for online machine translation. Companies and governments with confidentiality requirements often ban online translation or pay a premium to disable logging. To bring control back to the end user and demonstrate speed, we developed translateLocally. Running locally on a desktop or laptop CPU, translateLocally delivers cloud-like translation speed and quality even on 10 year old hardware. The open-source software is based on Marian and runs on Linux, Windows, and macOS.

Datasets: A Community Library for Natural Language Processing
Quentin Lhoest et al.

The scale, variety, and quantity of publicly-available NLP datasets has grown rapidly as researchers propose new tasks, larger models, and novel benchmarks. Datasets is a community library for contemporary NLP designed to support this ecosystem. Datasets aims to standardize end-user interfaces, versioning, and documentation, while providing a lightweight front-end that behaves similarly for small datasets as for internet-scale corpora. The design of the library incorporates a distributed, community-driven approach to adding datasets and documenting usage. After a year of development, the library now includes more than 650 unique datasets, has more than 250 contributors, and has helped support a variety of novel cross-dataset research projects and shared tasks. The library is available at https://github.com/huggingface/datasets.

MeetDot: Videoconferencing with Live Translation Captions
Arkady Arkhangorodsky et al.

We present MeetDot, a videoconferencing system with live translation captions overlaid on screen. The system aims to facilitate conversation between people who speak different languages, thereby reducing communication barriers between multilingual participants. Currently, our system supports speech and captions in 4 languages and combines automatic speech recognition (ASR) and machine translation (MT) in a cascade. We use the re-translation strategy to translate the streamed speech, resulting in caption flicker. Additionally, our system has very strict latency requirements to have acceptable call quality. We implement several features to enhance user experience and reduce their cognitive load, such as smooth scrolling captions and reducing caption flicker. The modular architecture allows us to integrate different ASR and MT services in our backend. Our system provides an integrated evaluation suite to optimize key intrinsic evaluation metrics such as accuracy, latency and erasure. Finally, we present an innovative cross-lingual word-guessing game as an extrinsic evaluation metric to measure end-to-end system performance. We plan to make our system open-source for research purposes.

T3-Vis: visual analytic for Training and fine-Tuning Transformers in NLP
Raymond Li et al.

Transformers are the dominant architecture in NLP, but their training and fine-tuning is still very challenging. In this paper, we present the design and implementation of a visual analytic framework for assisting researchers in such process, by providing them with valuable insights about the model's intrinsic properties and behaviours. Our framework offers an intuitive overview that allows the user to explore different facets of the model (e.g., hidden states, attention) through interactive visualization, and allows a suite of built-in algorithms that compute the importance of model components and different parts of the input sequence. Case studies and feedback from a user focus group indicate that the framework is useful, and suggest several improvements. Our framework is available at: https://github.com/raymondzmc/T3-Vis.

DomiKnowS: A Library for Integration of Symbolic Domain Knowledge in Deep Learning
Hossein Rajaby Faghihi et al.

We demonstrate a library for the integration of domain knowledge in deep learning architectures. Using this library, the structure of the data is expressed symbolically via graph declarations and the logical constraints over outputs or latent variables can be seamlessly added to the deep models. The domain knowledge can be defined explicitly, which improves the explainability of the models in addition to their performance and generalizability in the low-data regime. Several approaches for such integration of symbolic and sub-symbolic models have been introduced; however, there is no library to facilitate the programming for such integration in a generic way while various underlying algorithms can be used. Our library aims to simplify programming for such integration in both training and inference phases while separating the knowledge representation from learning algorithms. We showcase various NLP benchmark tasks and beyond. The framework is publicly available at Github (https://github.com/HLR/DomiKnowS).

IrEne-viz: Visualizing Energy Consumption of Transformer Models
Yash Kumar Lal et al.

IrEne is an energy prediction system that accurately predicts the interpretable inference energy consumption of a wide range of Transformer-based NLP models. We present the IrEne-viz tool, an online platform for visualizing and exploring energy consumption of various Transformer-based models easily. Additionally, we release a public API that can be used to access granular information about energy consumption of transformer models and their components. The live demo is available at http://stonybrooknlp.github.io/irene/demo/.

Project Debater APIs: Decomposing the AI Grand Challenge
Roy Bar-Haim et al.

Project Debater was revealed in 2019 as the first AI system that can debate human experts on complex topics. Engaging in a live debate requires a diverse set of skills, and Project Debater has been developed accordingly as a collection of components, each designed to perform a specific subtask. Project Debater APIs provide access to many of these capabilities, as well as to more recently developed ones. This diverse set of web services, publicly available for academic use, includes core NLP services, argument mining and analysis capabilities, and higher-level services for content summarization. We describe these APIs and their performance, and demonstrate how they can be used for building practical solutions. In particular, we will focus on Key Point Analysis, a novel technology that identifies the main points and their prevalence in a collection of texts such as survey responses and user reviews.

SeqAttack: On Adversarial Attacks for Named Entity Recognition
Walter Simoncini and Gerasimos Spanakis

Named Entity Recognition is a fundamental task in information extraction and is an essential element for various Natural Language Processing pipelines. Adversarial attacks have been shown to greatly affect the performance of text classification systems but knowledge about their effectiveness against named entity recognition models is limited. This paper investigates the effectiveness and portability of adversarial attacks from text classification to named entity recognition and the ability of adversarial training to counteract these attacks. We find that character-level and word-level attacks are the most effective, but adversarial training can grant significant protection at little to no expense of standard performance. Alongside our results, we also release SeqAttack, a framework to conduct adversarial attacks against token classification models (used in this work for named entity recognition) and a companion web application to inspect and cherry pick adversarial examples.

InVeRo-XL: Making Cross-Lingual Semantic Role Labeling Accessible with Intelligible Verbs and Roles
Simone Conia et al.

Notwithstanding the growing interest in cross-lingual techniques for Natural Language Processing, there has been a surprisingly small number of efforts aimed at the development of easy-to-use tools for cross-lingual Semantic Role Labeling. In this paper, we fill this gap and present InVeRo-XL, an off-the-shelf state-of-the-art system capable of annotating text with predicate sense and semantic role labels from 7 predicate-argument structure inventories in more than 40 languages. We hope that our system, with its easy-to-use RESTful API and Web interface, will become a valuable tool for the research community, encouraging the integration of sentence-level semantics into cross-lingual downstream tasks. InVeRo-XL is available online at http://nlp.uniroma1.it/invero.

TabPert: An Effective Platform for Tabular Perturbation
Nupur Jain et al.

To grasp the true reasoning ability, the Natural Language Inference model should be evaluated on counterfactual data. TabPert facilitates this by generation of such counterfactual data for assessing model tabular reasoning issues. TabPert allows the user to update a table, change the hypothesis, change the labels, and highlight rows that are important for hypothesis classification. TabPert also details the technique used to automatically produce the table, as well as the strategies employed to generate the challenging hypothesis. These counterfactual tables and hypotheses, as well as the metadata, is then used to explore the existing model's shortcomings methodically and quantitatively.

deepQuest-py: Large and Distilled Models for Quality Estimation
Fernando Alva-Manchego et al.

We introduce deepQuest-py, a framework for training and evaluation of large and light-weight models for Quality Estimation (QE). deepQuest-py provides access to (1) state-of-the-art models based on pre-trained Transformers for sentence-level and word-level QE; (2) light-weight and efficient sentence-level models implemented via knowledge distillation; and (3) a web interface for testing models and visualising their predictions. deepQuest-py is available at https://github.com/sheffieldnlp/deepQuest-py under a CC BY-NC-SA licence.

Social Event

Monday, November 8, 2021

The Social Event will be a Beach Party for everyone to mix and mingle and reacquaint themselves with those who have not been able to travel for the past (almost) two years. There will be dinner and entertainment and should be a lot of fun.

6

Main Conference: Tuesday, November 9

Overview

08:30 – 10:30 **Virtual Poster Session 3**
10:30 – 11:00 **Coffee Break**

Session 8

11:00 – 12:30						
	Machine Learning for NLP 5	Question Answering 3	Information Extraction 4	NLP Applications 2	Speech, Vision, Robotics, Multimodal Grounding 3	Semantics 4
	Bavaro 1	*Bavaro 2*	*Bavaro 3*	*Punta Cana 1,2*	*Punta Cana 3,4*	*Santo Domingo*
	[Posters&Demos]: In-person Poster 5					*Bavaro 4*

13:25 – 14:25 **SIGDAT Business Meeting**

Session 9

14:30 – 16:00						
	Machine Translation and Multilinguality 4	Interpretability and Analysis of Models for NLP 3	Information Extraction 5	Resources and Evaluation 4	Syntax	Efficient Methods for NLP 3
	Bavaro 1	*Bavaro 2*	*Bavaro 3*	*Punta Cana 1,2*	*Punta Cana 3,4*	*Santo Domingo*
	[Posters&Demos]: In-person Poster 6					*Bavaro 4*

16:00 – 16:30 **Coffee Break**
16:30 – 17:30 **Invited Talk 3 by Steven Bird** *Bavaro 1, 2, 3*
17:30 – 18:00 **Closing Ceremony**

Keynote Address: Steven Bird

LT4All!? Rethinking the Agenda

Tuesday, November 9, 2021, 16:30–17:30am

Bavaro 1, 2, 3

Abstract: The majority of the world's languages are oral, emergent, untranslatable, and tightly coupled to a place. Yet it seems that the agenda is to supply all languages with the technologies that have been developed for written languages. It is as though standardised writing were the optimal way to safeguard the future of any language. It is as though the function of a language is exclusively for transmitting information, and that the same information can be rendered into any language. It is as though we can capture and model language data independently of people, purpose, and place. What would it be like if language technologies respected the self-determination of a local speech community and supported aspirations concerning the local repertoire of speech varieties? The answer will be different in different places, but there may be value in taking a close look at an individual community and trying to discern broader themes. In this talk I will share from my experience of living and working in a remote Aboriginal community in the far north of Australia. Here, local people have been teaching me participatory, relational, strengths-based approaches that my students and I have been exploring in the design of language technologies. I will reflect on five years of personal experiences in this space and share thoughts concerning an agenda for language technology in the interest of minority speech communities, and hopes for creating a world that sustains its languages.

Biography: Steven Bird has spent 25 years pursuing scalable computational methods for capturing, enriching, and analysing data from endangered languages, drawing on fieldwork in West Africa, South America, and Melanesia. Over the past 5 years he has begun to work with remote Aboriginal communities in northern Australia. Steven has held academic positions at U Edinburgh, U Pennsylvania, UC Berkeley, and U Melbourne. He currently holds the positions of professor at Charles Darwin University, linguist at Nawarddeken Academy, and producer at languageparty.org.

Session 8 Overview – Tuesday, November 9, 2021

Track A	**Track B**	**Track C**	**Track D**	**Track E**	**Track F**	
Machine Learning for NLP 5	*Question Answering 3*	*Information Extraction 4*	*NLP Applications 2*	*Speech, Vision, Robotics, Multimodal Grounding 3*	*Semantics 4*	
Neuralizing Regular Expressions for Slot Filling *Jiang, Jin, and Tu*	Contrastive Domain Adaptation for Question Answering using Limited Text Corpora *Yue, Kratzwald, and Feuerriegel*	Set Generation Networks for End-to-End Knowledge Base Population *Sui et al.*	Exploring Methods for Generating Feedback Comments for Writing Learning *Hanawa, Nagata, and Inui*	Inflate and Shrink:Enriching and Reducing Interactions for Fast Text-Image Retrieval *Liu, Yu, and Li*	Integrating Deep Event-Level and Script-Level Information for Script Event Prediction *Bai et al.*	11:00
Causal Direction of Data Collection Matters: Implications of Causal and Anticausal Learning for NLP *Jin et al.*	Case-based Reasoning for Natural Language Queries over Knowledge Bases *Das et al.*	Knowing False Negatives: An Adversarial Training Method for Distantly Supervised Relation Extraction *Hao, Yu, and Hu*	A Role-Selected Sharing Network for Joint Machine-Human Chatting Handoff and Service Satisfaction Analysis *Liu et al.*	On Pursuit of Designing Multi-modal Transformer for Video Grounding *Cao et al.*	Decoding Word Embeddings with Brain-Based Semantic Features *Chersoni et al.*	11:15
Raise a Child in Large Language Model: Towards Effective and Generalizable Fine-tuning *Xu et al.*	Distantly-Supervised Dense Retrieval Enables Open-Domain Question Answering without Evidence Annotation *Zhao et al.*	Progressive Adversarial Learning for Bootstrapping: A Case Study on Entity Set Expansion *Yan, Han, and Sun*	Meta Distant Transfer Learning for Pre-trained Language Models *Wang et al.*	COVR: A Test-Bed for Visually Grounded Compositional Generalization with Real Images *Bogin et al.*	Syntax Role for Neural Semantic Role Labeling *Li et al.*	11:30
Knowledge Graph Representation Learning using Ordinary Differential Equations *Nayyeri et al.*	PAQ: 65 Million Probably-Asked Questions and What You Can Do With Them *Lewis et al.*	Uncovering Main Causalities for Long-tailed Information Extraction *Nan et al.*	UniKER: A Unified Framework for Combining Embedding and Definite Horn Rule Reasoning for Knowledge Graph Inference *Zhang and Sun*	Vision-and-Language or Vision-for-Language? On Cross-Modal Influence in Multimodal Transformers *Frank, Bugliarello, and Elliott*	QA-Align: Representing Cross-Text Content Overlap by Aligning Question-Answer Propositions *Brook Weiss et al.*	11:45
KnowMAN: Weakly Supervised Multinomial Adversarial Networks *März et al.*	What's in a Name? Answer Equivalence For Open-Domain Question Answering *Si, Zhao, and Boyd-Graber*	Maximal Clique Based Non-Autoregressive Open Information Extraction *Yu et al.*	Wasserstein Selective Transfer Learning for Cross-domain Text Mining *Feng et al.*	HypMix: Hyperbolic Interpolative Data Augmentation *Sawhney et al.*	PICARD: Parsing Incrementally for Constrained Auto-Regressive Decoding from Language Models *Scholak, Schucher, and Bahdanau*	12:00

	Track A	**Track B**	**Track C**	**Track D**	**Track E**	**Track F**
	Machine Learning for NLP 5	*Question Answering 3*	*Information Extraction 4*	*NLP Applications 2*	*Speech, Vision, Robotics, Multimodal Grounding 3*	*Semantics 4*
12:10	ONION: A Simple and Effective Defense Against Textual Backdoor Attacks *Qi et al.*	Evaluation Paradigms in Question Answering *Rodriguez and Boyd-Graber*	A Relation-Oriented Clustering Method for Open Relation Extraction *Zhao et al.*	Jointly Learning to Repair Code and Generate Commit Message *Bai et al.*	Multimodal Pretraining Unmasked: A Meta-Analysis and a Unified Framework of Vision-and-Language BERTs *Bugliarello et al.*	Exploiting Twitter as Source of Large Corpora of Weakly Similar Pairs for Semantic Sentence Embeddings *Di Giovanni and Brambilla*
12:20	Value-aware Approximate Attention *Gupta and Berant*	Numerical reasoning in machine reading comprehension tasks: are we there yet? *Al-Negheimish, Madhyastha, and Russo*				Guilt by Association: Emotion Intensities in Lexical Representations *Raji and Melo*

Parallel Session 8

Session 8A: Machine Learning for NLP 5

Bavaro 1 *Chair: Carolin Lawrence*

Neuralizing Regular Expressions for Slot Filling

Chengyue Jiang, Zijian Jin, and Kewei Tu 11:00–11:15

Neural models and symbolic rules such as regular expressions have their respective merits and weaknesses. In this paper, we study the integration of the two approaches for the slot filling task by converting regular expressions into neural networks. Specifically, we first convert regular expressions into a special form of finite-state transducers, then unfold its approximate inference algorithm as a bidirectional recurrent neural model that performs slot filling via sequence labeling. Experimental results show that our model has superior zero-shot and few-shot performance and stays competitive when there are sufficient training data.

Causal Direction of Data Collection Matters: Implications of Causal and Anticausal Learning for NLP

Zhijing Jin et al. 11:15–11:30

The principle of independent causal mechanisms (ICM) states that generative processes of real world data consist of independent modules which do not influence or inform each other. While this idea has led to fruitful developments in the field of causal inference, it is not widely-known in the NLP community. In this work, we argue that the causal direction of the data collection process bears nontrivial implications that can explain a number of published NLP findings, such as differences in semi-supervised learning (SSL) and domain adaptation (DA) performance across different settings. We categorize common NLP tasks according to their causal direction and empirically assay the validity of the ICM principle for text data using minimum description length. We conduct an extensive meta-analysis of over 100 published SSL and 30 DA studies, and find that the results are consistent with our expectations based on causal insights. This work presents the first attempt to analyze the ICM principle in NLP, and provides constructive suggestions for future modeling choices.

Raise a Child in Large Language Model: Towards Effective and Generalizable Fine-tuning

Runxin Xu et al. 11:30–11:45

Recent pretrained language models extend from millions to billions of parameters. Thus the need to fine-tune an extremely large pretrained model with a limited training corpus arises in various downstream tasks. In this paper, we propose a straightforward yet effective fine-tuning technique, Child-Tuning, which updates a subset of parameters (called child network) of large pretrained models via strategically masking out the gradients of the non-child network during the backward process. Experiments on various downstream tasks in GLUE benchmark show that Child-Tuning consistently outperforms the vanilla fine-tuning by 1.5~8.6 average score among four different pretrained models, and surpasses the prior fine-tuning techniques by 0.6~1.3 points. Furthermore, empirical results on domain transfer and task transfer show that Child-Tuning can obtain better generalization performance by large margins.

Knowledge Graph Representation Learning using Ordinary Differential Equations

Mojtaba Nayyeri et al. 11:45–12:00

Knowledge Graph Embeddings (KGEs) have shown promising performance on link prediction tasks by mapping the entities and relations from a knowledge graph into a geometric space. The capability of KGEs in preserving graph characteristics including structural aspects and semantics, highly depends on the design of their score function, as well as the inherited abilities from the underlying geometry. Many KGEs use the Euclidean geometry which renders them incapable of preserving complex structures and consequently causes wrong inferences by the models. To address this problem, we propose a neuro differential KGE that embeds nodes of a KG on the trajectories of Ordinary Differential Equations (ODEs). To this end, we represent each relation (edge) in a KG as a vector field on several manifolds. We specifically parameterize ODEs by a neural network to represent complex manifolds and complex vector fields on the manifolds. Therefore, the underlying embedding space is capable to assume the shape of various geometric forms to encode heterogeneous subgraphs. Experiments on synthetic and benchmark datasets using state-of-the-art KGE models justify the ODE trajectories as a means to enable structure preservation and consequently avoiding wrong inferences.

KnowMAN: Weakly Supervised Multinomial Adversarial Networks

Luisa März et al. 12:00–12:10

The absence of labeled data for training neural models is often addressed by leveraging knowledge about the specific task, resulting in heuristic but noisy labels. The knowledge is captured in labeling functions, which detect certain regularities or patterns in the training samples and annotate corresponding labels for training. This process of weakly supervised training may result in an over-reliance on the signals captured by the labeling functions and hinder models to exploit other signals or to generalize well. We propose KnowMAN, an adversarial scheme that enables to control influence of signals associated with specific labeling functions. KnowMAN forces the network to learn representations that are invariant to those signals and to pick up other signals that are more generally associated with an output label. KnowMAN strongly improves results compared to direct weakly supervised learning with a pre-trained transformer language model and a feature-based

baseline.

ONION: A Simple and Effective Defense Against Textual Backdoor Attacks
Fanchao Qi et al. 12:10–12:20

Backdoor attacks are a kind of emergent training-time threat to deep neural networks (DNNs). They can manipulate the output of DNNs and possess high insidiousness. In the field of natural language processing, some attack methods have been proposed and achieve very high attack success rates on multiple popular models. Nevertheless, there are few studies on defending against textual backdoor attacks. In this paper, we propose a simple and effective textual backdoor defense named ONION, which is based on outlier word detection and, to the best of our knowledge, is the first method that can handle all the textual backdoor attack situations. Experiments demonstrate the effectiveness of our model in defending BiLSTM and BERT against five different backdoor attacks. All the code and data of this paper can be obtained at https://github.com/thunlp/ONION.

Value-aware Approximate Attention
Ankit Gupta and Jonathan Berant 12:20–12:30

Following the success of dot-product attention in Transformers, numerous approximations have been recently proposed to address its quadratic complexity with respect to the input length. However, all approximations thus far have ignored the contribution of the *value vectors* to the quality of approximation. In this work, we argue that research efforts should be directed towards approximating the true output of the attention sub-layer, which includes the value vectors. We propose a value-aware objective, and show theoretically and empirically that an optimal approximation of a value-aware objective substantially outperforms an optimal approximation that ignores values, in the context of language modeling. Moreover, we show that the choice of kernel function for computing attention similarity can substantially affect the quality of sparse approximations, where kernel functions that are less skewed are more affected by the value vectors.

Session 8B: Question Answering 3

Bavaro 2 *Chair: Jonathan Berant*

Contrastive Domain Adaptation for Question Answering using Limited Text Corpora

Zhenrui Yue, Bernhard Kratzwald, and Stefan Feuerriegel 11:00–11:15

Question generation has recently shown impressive results in customizing question answering (QA) systems to new domains. These approaches circumvent the need for manually annotated training data from the new domain and, instead, generate synthetic question-answer pairs that are used for training. However, existing methods for question generation rely on large amounts of synthetically generated datasets and costly computational resources, which render these techniques widely inaccessible when the text corpora is of limited size. This is problematic as many niche domains rely on small text corpora, which naturally restricts the amount of synthetic data that can be generated. In this paper, we propose a novel framework for domain adaptation called contrastive domain adaptation for QA (CAQA). Specifically, CAQA combines techniques from question generation and domain-invariant learning to answer out-of-domain questions in settings with limited text corpora. Here, we train a QA system on both source data and generated data from the target domain with a contrastive adaptation loss that is incorporated in the training objective. By combining techniques from question generation and domain-invariant learning, our model achieved considerable improvements compared to state-of-the-art baselines.

Case-based Reasoning for Natural Language Queries over Knowledge Bases

Rajarshi Das et al. 11:15–11:30

It is often challenging to solve a complex problem from scratch, but much easier if we can access other similar problems with their solutions — a paradigm known as case-based reasoning (CBR). We propose a neuro-symbolic CBR approach (CBR-KBQA) for question answering over large knowledge bases. CBR-KBQA consists of a nonparametric memory that stores cases (question and logical forms) and a parametric model that can generate a logical form for a new question by retrieving cases that are relevant to it. On several KBQA datasets that contain complex questions, CBR-KBQA achieves competitive performance. For example, on the CWQ dataset, CBR-KBQA outperforms the current state of the art by 11% on accuracy. Furthermore, we show that CBR-KBQA is capable of using new cases *without* any further training: by incorporating a few human-labeled examples in the case memory, CBR-KBQA is able to successfully generate logical forms containing unseen KB entities as well as relations.

Distantly-Supervised Dense Retrieval Enables Open-Domain Question Answering without Evidence Annotation

Chen Zhao et al. 11:30–11:45

Open-domain question answering answers a question based on evidence retrieved from a large corpus. State-of-the-art neural approaches require intermediate evidence annotations for training. However, such intermediate annotations are expensive, and methods that rely on them cannot transfer to the more common setting, where only question—answer pairs are available. This paper investigates whether models can learn to find evidence from a large corpus, with only distant supervision from answer labels for model training, thereby generating no additional annotation cost. We introduce a novel approach (DistDR) that iteratively improves over a weak retriever by alternately finding evidence from the up-to-date model and encouraging the model to learn the most likely evidence. Without using any evidence labels, DistDR is on par with fully-supervised state-of-the-art methods on both multi-hop and single-hop QA benchmarks. Our analysis confirms that DistDR finds more accurate evidence over iterations, which leads to model improvements. The code is available at https://github.com/henryzhao5852/DistDR.

PAQ: 65 Million Probably-Asked Questions and What You Can Do With Them

Patrick Lewis et al. 11:45–12:00

Open-domain Question Answering models which directly leverage question-answer (QA) pairs, such as closed-book QA (CBQA) models and QA-pair retrievers, show promise in terms of speed and memory compared to conventional models which retrieve and read from text corpora. QA-pair retrievers also offer interpretable answers, a high degree of control, and are trivial to update at test time with new knowledge. However, these models fall short of the accuracy of retrieve-and-read systems, as substantially less knowledge is covered by the available QA-pairs relative to text corpora like Wikipedia. To facilitate improved QA-pair models, we introduce Probably Asked Questions (PAQ), a very large resource of 65M automatically-generated QA-pairs. We introduce a new QA-pair retriever, RePAQ, to complement PAQ. We find that PAQ pre-empts and caches test questions, enabling RePAQ to match the accuracy of recent retrieve-and-read models, whilst being significantly faster. Using PAQ, we train CBQA models which outperform comparable baselines by 5%, but trail RePAQ by over 15%, indicating the effectiveness of explicit retrieval. RePAQ can be configured for size (under 500MB) or speed (over 1K questions per second) whilst retaining high accuracy. Lastly, we demonstrate RePAQ's strength at selective QA, abstaining from answering when it is likely to be incorrect. This enables RePAQ to "back-off" to a more expensive state-of-the-art model, leading to a combined system which is both more accurate and 2x faster than the state-of-the-art model alone.

What's in a Name? Answer Equivalence For Open-Domain Question Answering

Chenglei Si, Chen Zhao, and Jordan Boyd-Graber 12:00–12:10

A flaw in QA evaluation is that annotations often only provide one gold answer. Thus, model predictions semantically equivalent to the answer but superficially different are considered incorrect. This work explores

mining alias entities from knowledge bases and using them as additional gold answers (i.e., equivalent answers). We incorporate answers for two settings: evaluation with additional answers and model training with equivalent answers. We analyse three QA benchmarks: Natural Questions, TriviaQA, and SQuAD. Answer expansion increases the exact match score on all datasets for evaluation, while incorporating it helps model training over real-world datasets. We ensure the additional answers are valid through a human post hoc evaluation.

Evaluation Paradigms in Question Answering
Pedro Rodriguez and Jordan Boyd-Graber 12:10–12:20

Question answering (QA) primarily descends from two branches of research: (1) Alan Turing's investigation of machine intelligence at Manchester University and (2) Cyril Cleverdon's comparison of library card catalog indices at Cranfield University. This position paper names and distinguishes these paradigms. Despite substantial overlap, subtle but significant distinctions exert an outsize influence on research. While one evaluation paradigm values creating more intelligent QA systems, the other paradigm values building QA systems that appeal to users. By better understanding the epistemic heritage of QA, researchers, academia, and industry can more effectively accelerate QA research.

Numerical reasoning in machine reading comprehension tasks: are we there yet?
Hadeel Al-Negheimish, Pranava Madhyastha, and Alessandra Russo 12:20–12:30

Numerical reasoning based machine reading comprehension is a task that involves reading comprehension along with using arithmetic operations such as addition, subtraction, sorting and counting. The DROP benchmark (Dua et al., 2019) is a recent dataset that has inspired the design of NLP models aimed at solving this task. The current standings of these models in the DROP leaderboard, over standard metrics, suggests that the models have achieved near-human performance. However, does this mean that these models have learned to reason? In this paper, we present a controlled study on some of the top-performing model architectures for the task of numerical reasoning. Our observations suggest that the standard metrics are incapable of measuring progress towards such tasks.

Session 8C: Information Extraction 4

Bavaro 3 *Chair: Christos Christodoulopoulos*

Set Generation Networks for End-to-End Knowledge Base Population

Dianbo Sui et al. 11:00–11:15

The task of knowledge base population (KBP) aims to discover facts about entities from texts and expand a knowledge base with these facts. Previous studies shape end-to-end KBP as a machine translation task, which is required to convert unordered fact into a sequence according to a pre-specified order. However, the facts stated in a sentence are unordered in essence. In this paper, we formulate end-to-end KBP as a direct set generation problem, avoiding considering the order of multiple facts. To solve the set generation problem, we propose networks featured by transformers with non-autoregressive parallel decoding. Unlike previous approaches that use an autoregressive decoder to generate facts one by one, the proposed networks can directly output the final set of facts in one shot. Furthermore, to train the networks, we also design a set-based loss that forces unique predictions via bipartite matching. Compared with cross-entropy loss that highly penalizes small shifts in fact order, the proposed bipartite matching loss is invariant to any permutation of predictions. Benefiting from getting rid of the burden of predicting the order of multiple facts, our proposed networks achieve state-of-the-art (SoTA) performance on two benchmark datasets.

Knowing False Negatives: An Adversarial Training Method for Distantly Supervised Relation Extraction

Kailong Hao, Botao Yu, and Wei Hu 11:15–11:30

Distantly supervised relation extraction (RE) automatically aligns unstructured text with relation instances in a knowledge base (KB). Due to the incompleteness of current KBs, sentences implying certain relations may be annotated as N/A instances, which causes the so-called false negative (FN) problem. Current RE methods usually overlook this problem, inducing improper biases in both training and testing procedures. To address this issue, we propose a two-stage approach. First, it finds out possible FN samples by heuristically leveraging the memory mechanism of deep neural networks. Then, it aligns those unlabeled data with the training data into a unified feature space by adversarial training to assign pseudo labels and further utilize the information contained in them. Experiments on two wildly-used benchmark datasets demonstrate the effectiveness of our approach.

Progressive Adversarial Learning for Bootstrapping: A Case Study on Entity Set Expansion

Lingyong Yan, Xianpei Han, and Le Sun 11:30–11:45

Bootstrapping has become the mainstream method for entity set expansion. Conventional bootstrapping methods mostly define the expansion boundary using seed-based distance metrics, which heavily depend on the quality of selected seeds and are hard to be adjusted due to the extremely sparse supervision. In this paper, we propose BootstrapGAN, a new learning method for bootstrapping which jointly models the bootstrapping process and the boundary learning process in a GAN framework. Specifically, the expansion boundaries of different bootstrapping iterations are learned via different discriminator networks; the bootstrapping network is the generator to generate new positive entities, and the discriminator networks identify the expansion boundaries by trying to distinguish the generated entities from known positive entities. By iteratively performing the above adversarial learning, the generator and the discriminators can reinforce each other and be progressively refined along the whole bootstrapping process. Experiments show that BootstrapGAN achieves the new state-of-the-art entity set expansion performance.

Uncovering Main Causalities for Long-tailed Information Extraction

Guoshun Nan et al. 11:45–12:00

Information Extraction (IE) aims to extract structural information from unstructured texts. In practice, long-tailed distributions caused by the selection bias of a dataset may lead to incorrect correlations, also known as spurious correlations, between entities and labels in the conventional likelihood models. This motivates us to propose counterfactual IE (CFIE), a novel framework that aims to uncover the main causalities behind data in the view of causal inference. Specifically, 1) we first introduce a unified structural causal model (SCM) for various IE tasks, describing the relationships among variables; 2) with our SCM, we then generate counterfactuals based on an explicit language structure to better calculate the direct causal effect during the inference stage; 3) we further propose a novel debiasing approach to yield more robust predictions. Experiments on three IE tasks across five public datasets show the effectiveness of our CFIE model in mitigating the spurious correlation issues.

Maximal Clique Based Non-Autoregressive Open Information Extraction

Bowen Yu et al. 12:00–12:15

Open Information Extraction (OpenIE) aims to discover textual facts from a given sentence. In essence, the facts contained in plain text are unordered. However, the popular OpenIE systems usually output facts sequentially in the way of predicting the next fact conditioned on the previous decoded ones, which enforce an unnecessary order on the facts and involve the error accumulation between autoregressive steps. To break this bottleneck, we propose MacroIE, a novel non-autoregressive framework for OpenIE. MacroIE firstly constructs a fact graph based on the table filling scheme, in which each node denotes a fact element, and an edge links two nodes that belong to the same fact. Then OpenIE can be reformulated as a non-parametric process

of finding maximal cliques from the graph. It directly outputs the final set of facts in one go, thus getting rid of the burden of predicting fact order, as well as the error propagation between facts. Experiments conducted on two benchmark datasets show that our proposed model significantly outperforms current state-of-the-art methods, beats the previous systems by as much as 5.7 absolute gain in F1 score.

A Relation-Oriented Clustering Method for Open Relation Extraction
Jun Zhao et al. 12:15–12:30

The clustering-based unsupervised relation discovery method has gradually become one of the important methods of open relation extraction (OpenRE). However, high-dimensional vectors can encode complex linguistic information which leads to the problem that the derived clusters cannot explicitly align with the relational semantic classes. In this work, we propose a relation-oriented clustering model and use it to identify the novel relations in the unlabeled data. Specifically, to enable the model to learn to cluster relational data, our method leverages the readily available labeled data of pre-defined relations to learn a relation-oriented representation. We minimize distance between the instance with same relation by gathering the instances towards their corresponding relation centroids to form a cluster structure, so that the learned representation is cluster-friendly. To reduce the clustering bias on predefined classes, we optimize the model by minimizing a joint objective on both labeled and unlabeled data. Experimental results show that our method reduces the error rate by 29.2% and 15.7%, on two datasets respectively, compared with current SOTA methods.

Session 8D: NLP Applications 2

Punta Cana 1,2 *Chair: Emilia Apostolova*

Exploring Methods for Generating Feedback Comments for Writing Learning

Kazuaki Hanawa, Ryo Nagata, and Kentaro Inui 11:00–11:15

The task of generating explanatory notes for language learners is known as feedback comment generation. Although various generation techniques are available, little is known about which methods are appropriate for this task. Nagata (2019) demonstrates the effectiveness of neural-retrieval-based methods in generating feedback comments for preposition use. Retrieval-based methods have limitations in that they can only output feedback comments existing in a given training data. Furthermore, feedback comments can be made on other grammatical and writing items than preposition use, which is still unaddressed. To shed light on these points, we investigate a wider range of methods for generating many feedback comments in this study. Our close analysis of the type of task leads us to investigate three different architectures for comment generation: (i) a neural-retrieval-based method as a baseline, (ii) a pointer-generator-based generation method as a neural seq2seq method, (iii) a retrieve-and-edit method, a hybrid of (i) and (ii). Intuitively, the pointer-generator should outperform neural-retrieval, and retrieve-and-edit should perform best. However, in our experiments, this expectation is completely overturned. We closely analyze the results to reveal the major causes of these counter-intuitive results and report on our findings from the experiments.

A Role-Selected Sharing Network for Joint Machine-Human Chatting Handoff and Service Satisfaction Analysis

Jiawei Liu et al. 11:15–11:30

Chatbot is increasingly thriving in different domains, however, because of unexpected discourse complexity and training data sparseness, its potential distrust hatches vital apprehension. Recently, Machine-Human Chatting Handoff (MHCH), predicting chatbot failure and enabling human-algorithm collaboration to enhance chatbot quality, has attracted increasing attention from industry and academia. In this study, we propose a novel model, Role-Selected Sharing Network (RSSN), which integrates both dialogue satisfaction estimation and handoff prediction in one multi-task learning framework. Unlike prior efforts in dialog mining, by utilizing local user satisfaction as a bridge, global satisfaction detector and handoff predictor can effectively exchange critical information. Specifically, we decouple the relation and interaction between the two tasks by the role information after the shared encoder. Extensive experiments on two public datasets demonstrate the effectiveness of our model.

Meta Distant Transfer Learning for Pre-trained Language Models

Chengyu Wang et al. 11:30–11:45

With the wide availability of Pre-trained Language Models (PLMs), multi-task fine-tuning across domains has been extensively applied. For tasks related to distant domains with different class label sets, PLMs may memorize non-transferable knowledge for the target domain and suffer from negative transfer. Inspired by meta-learning, we propose the Meta Distant Transfer Learning (Meta-DTL) framework to learn the cross-task knowledge for PLM-based methods. Meta-DTL first employs task representation learning to mine implicit relations among multiple tasks and classes. Based on the results, it trains a PLM-based meta-learner to capture the transferable knowledge across tasks. The weighted maximum entropy regularizers are proposed to make meta-learner more task-agnostic and unbiased. Finally, the meta-learner can be fine-tuned to fit each task with better parameter initialization. We evaluate Meta-DTL using both BERT and ALBERT on seven public datasets. Experiment results confirm the superiority of Meta-DTL as it consistently outperforms strong baselines. We find that Meta-DTL is highly effective when very few data is available for the target task.

UniKER: A Unified Framework for Combining Embedding and Definite Horn Rule Reasoning for Knowledge Graph Inference

Ming Zhang and Yizhou Sun 11:45–12:00

Knowledge graph inference has been studied extensively due to its wide applications. It has been addressed by two lines of research, i.e., the more traditional logical rule reasoning and the more recent knowledge graph embedding (KGE). Several attempts have been made to combine KGE and logical rules for better knowledge graph inference. Unfortunately, they either simply treat logical rules as additional constraints into KGE loss or use probabilistic model to approximate the exact logical inference (i.e., MAX-SAT). Even worse, both approaches need to sample ground rules to tackle the scalability issue, as the total number of ground rules is intractable in practice, making them less effective in handling logical rules. In this paper, we propose a novel framework UniKER to address these challenges by restricting logical rules to be definite Horn rules, which can fully exploit the knowledge in logical rules and enable the mutual enhancement of logical rule-based reasoning and KGE in an extremely efficient way. Extensive experiments have demonstrated that our approach is superior to existing state-of-the-art algorithms in terms of both efficiency and effectiveness.

Wasserstein Selective Transfer Learning for Cross-domain Text Mining

Lingyun Feng et al. 12:00–12:15

Transfer learning (TL) seeks to improve the learning of a data-scarce target domain by using information from source domains. However, the source and target domains usually have different data distributions, which may lead to negative transfer. To alleviate this issue, we propose a Wasserstein Selective Transfer Learning (WSTL) method. Specifically, the proposed method considers a reinforced selector to select helpful data for transfer

learning. We further use a Wasserstein-based discriminator to maximize the empirical distance between the selected source data and target data. The TL module is then trained to minimize the estimated Wasserstein distance in an adversarial manner and provides domain invariant features for the reinforced selector. We adopt an evaluation metric based on the performance of the TL module as delayed reward and a Wasserstein-based metric as immediate rewards to guide the reinforced selector learning. Compared with the competing TL approaches, the proposed method selects data samples that are closer to the target domain. It also provides better state features and reward signals that lead to better performance with faster convergence. Extensive experiments on three real-world text mining tasks demonstrate the effectiveness of the proposed method.

Jointly Learning to Repair Code and Generate Commit Message
Jiaqi Bai et al. 12:15–12:30

We propose a novel task of jointly repairing program codes and generating commit messages. Code repair and commit message generation are two essential and related tasks for software development. However, existing work usually performs the two tasks independently. We construct a multilingual triple dataset including buggy code, fixed code, and commit messages for this novel task. We first introduce a cascaded method with two models, one is to generate the fixed code first, and the other generates the commit message based on the fixed and original codes. We enhance the cascaded method with different training approaches, including the teacher-student method, the multi-task method, and the back-translation method. To deal with the error propagation problem of the cascaded method, we also propose a joint model that can both repair the program code and generate the commit message in a unified framework. Massive experiments on our constructed buggy-fixed-commit dataset reflect the challenge of this task and that the enhanced cascaded model and the proposed joint model significantly outperform baselines in both quality of code and commit messages.

Session 8E: Speech, Vision, Robotics, Multimodal Grounding 3

Punta Cana 3,4 *Chair: Hung-yi Lee*

Inflate and Shrink:Enriching and Reducing Interactions for Fast Text-Image Retrieval

Haoliang Liu, Tan Yu, and Ping Li 11:00–11:15

By exploiting the cross-modal attention, cross-BERT methods have achieved state-of-the-art accuracy in cross-modal retrieval. Nevertheless, the heavy text-image interactions in the cross-BERT model are prohibitively slow for large-scale retrieval. Late-interaction methods trade off retrieval accuracy and efficiency by exploiting cross-modal interaction only in the late stage, attaining a satisfactory retrieval speed. In this work, we propose an inflating and shrinking approach to further boost the efficiency and accuracy of late-interaction methods. The inflating operation plugs several codes in the input of the encoder to exploit the text-image interactions more thoroughly for higher retrieval accuracy. Then the shrinking operation gradually reduces the text-image interactions through knowledge distilling for higher efficiency. Through an inflating operation followed by a shrinking operation, both efficiency and accuracy of a late-interaction model are boosted. Systematic experiments on public benchmarks demonstrate the effectiveness of our inflating and shrinking approach.

On Pursuit of Designing Multi-modal Transformer for Video Grounding

Meng Cao et al. 11:15–11:30

Video grounding aims to localize the temporal segment corresponding to a sentence query from an untrimmed video. Almost all existing video grounding methods fall into two frameworks: 1) Top-down model: It predefines a set of segment candidates and then conducts segment classification and regression. 2) Bottom-up model: It directly predicts frame-wise probabilities of the referential segment boundaries. However, all these methods are not end-to-end, i.e., they always rely on some time-consuming post-processing steps to refine predictions. To this end, we reformulate video grounding as a set prediction task and propose a novel end-to-end multi-modal Transformer model, dubbed as GTR. Specifically, GTR has two encoders for video and language encoding, and a cross-modal decoder for grounding prediction. To facilitate the end-to-end training, we use a Cubic Embedding layer to transform the raw videos into a set of visual tokens. To better fuse these two modalities in the decoder, we design a new Multi-head Cross-Modal Attention. The whole GTR is optimized via a Many-to-One matching loss. Furthermore, we conduct comprehensive studies to investigate different model design choices. Extensive results on three benchmarks have validated the superiority of GTR. All three typical GTR variants achieve record-breaking performance on all datasets and metrics, with several times faster inference speed.

COVR: A Test-Bed for Visually Grounded Compositional Generalization with Real Images

Ben Bogin et al. 11:30–11:45

While interest in models that generalize at test time to new compositions has risen in recent years, benchmarks in the visually-grounded domain have thus far been restricted to synthetic images. In this work, we propose COVR, a new test-bed for visually-grounded compositional generalization with real images. To create COVR, we use real images annotated with scene graphs, and propose an almost fully automatic procedure for generating question-answer pairs along with a set of context images. COVR focuses on questions that require complex reasoning, including higher-order operations such as quantification and aggregation. Due to the automatic generation process, COVR facilitates the creation of compositional splits, where models at test time need to generalize to new concepts and compositions in a zero- or few-shot setting. We construct compositional splits using COVR and demonstrate a myriad of cases where state-of-the-art pre-trained language-and-vision models struggle to compositionally generalize.

Vision-and-Language or Vision-for-Language? On Cross-Modal Influence in Multimodal Transformers

Stella Frank, Emanuele Bugliarello, and Desmond Elliott 11:45–12:00

Pretrained vision-and-language BERTs aim to learn representations that combine information from both modalities. We propose a diagnostic method based on cross-modal input ablation to assess the extent to which these models actually integrate cross-modal information. This method involves ablating inputs from one modality, either entirely or selectively based on cross-modal grounding alignments, and evaluating the model prediction performance on the other modality. Model performance is measured by modality-specific tasks that mirror the model pretraining objectives (e.g. masked language modelling for text). Models that have learned to construct cross-modal representations using both modalities are expected to perform worse when inputs are missing from a modality. We find that recently proposed models have much greater relative difficulty predicting text when visual information is ablated, compared to predicting visual object categories when text is ablated, indicating that these models are not symmetrically cross-modal.

HypMix: Hyperbolic Interpolative Data Augmentation

Ramit Sawhney et al. 12:00–12:15

Interpolation-based regularisation methods for data augmentation have proven to be effective for various tasks and modalities. These methods involve performing mathematical operations over the raw input samples or their latent states representations - vectors that often possess complex hierarchical geometries. However, these operations are performed in the Euclidean space, simplifying these representations, which may lead to distorted and noisy interpolations. We propose HypMix, a novel model-, data-, and modality-agnostic interpolative data

augmentation technique operating in the hyperbolic space, which captures the complex geometry of input and hidden state hierarchies better than its contemporaries. We evaluate HypMix on benchmark and low resource datasets across speech, text, and vision modalities, showing that HypMix consistently outperforms state-of-the-art data augmentation techniques. In addition, we demonstrate the use of HypMix in semi-supervised settings. We further probe into the adversarial robustness and qualitative inferences we draw from HypMix that elucidate the efficacy of the Riemannian hyperbolic manifolds for interpolation-based data augmentation.

Multimodal Pretraining Unmasked: A Meta-Analysis and a Unified Framework of Vision-and-Language BERTs
Emanuele Bugliarello et al. 12:15–12:30

Large-scale pretraining and task-specific fine-tuning is now the standard methodology for many tasks in computer vision and natural language processing. Recently, a multitude of methods have been proposed for pretraining vision and language BERTs to tackle challenges at the intersection of these two key areas of AI. These models can be categorised into either single-stream or dual-stream encoders. We study the differences between these two categories, and show how they can be unified under a single theoretical framework. We then conduct controlled experiments to discern the empirical differences between five VL BERTs. Our experiments show that training data and hyperparameters are responsible for most of the differences between the reported results, but they also reveal that the embedding layer plays a crucial role in these massive models.

Session 8F: Semantics 4

Santo Domingo *Chair: James Henderson*

Integrating Deep Event-Level and Script-Level Information for Script Event Prediction

Long Bai et al. 11:00–11:15

Scripts are structured sequences of events together with the participants, which are extracted from the texts. Script event prediction aims to predict the subsequent event given the historical events in the script. Two kinds of information facilitate this task, namely, the event-level information and the script-level information. At the event level, existing studies view an event as a verb with its participants, while neglecting other useful properties, such as the state of the participants. At the script level, most existing studies only consider a single event sequence corresponding to one common protagonist. In this paper, we propose a Transformer-based model, called MCPredictor, which integrates deep event-level and script-level information for script event prediction. At the event level, MCPredictor utilizes the rich information in the text to obtain more comprehensive event semantic representations. At the script-level, it considers multiple event sequences corresponding to different participants of the subsequent event. The experimental results on the widely-used New York Times corpus demonstrate the effectiveness and superiority of the proposed model.

Decoding Word Embeddings with Brain-Based Semantic Features

Emmanuele Chersoni et al. 11:15–11:30

Word embeddings are vectorial semantic representations built with either counting or predicting techniques aimed at capturing shades of meaning from word co-occurrences. Since their introduction, these representations have been criticized for lacking interpretable dimensions. This property of word embeddings limits our understanding of the semantic features they actually encode. Moreover, it contributes to the "black box" nature of the tasks in which they are used, since the reasons for word embedding performance often remain opaque to humans. In this contribution, we explore the semantic properties encoded in word embeddings by mapping them onto interpretable vectors, consisting of explicit and neurobiologically motivated semantic features (Binder et al. 2016). Our exploration takes into account different types of embeddings, including factorized count vectors and predict models (Skip-Gram, GloVe, etc.), as well as the most recent contextualized representations (i.e., ELMo and BERT). In our analysis, we first evaluate the quality of the mapping in a retrieval task, then we shed light on the semantic features that are better encoded in each embedding type. A large number of probing tasks is finally set to assess how the original and the mapped embeddings perform in discriminating semantic categories. For each probing task, we identify the most relevant semantic features and we show that there is a correlation between the embedding performance and how they encode those features. This study sets itself as a step forward in understanding which aspects of meaning are captured by vector spaces, by proposing a new and simple method to carve human-interpretable semantic representations from distributional vectors.

Syntax Role for Neural Semantic Role Labeling

Zuchao Li et al. 11:30–11:45

Semantic role labeling (SRL) is dedicated to recognizing the semantic predicate-argument structure of a sentence. Previous studies in terms of traditional models have shown syntactic information can make remarkable contributions to SRL performance; however, the necessity of syntactic information was challenged by a few recent neural SRL studies that demonstrate impressive performance without syntactic backbones and suggest that syntax information becomes much less important for neural semantic role labeling, especially when paired with recent deep neural network and large-scale pre-trained language models. Despite this notion, the neural SRL field still lacks a systematic and full investigation on the relevance of syntactic information in SRL, for both dependency and both monolingual and multilingual settings. This paper intends to quantify the importance of syntactic information for neural SRL in the deep learning framework. We introduce three typical SRL frameworks (baselines), sequence-based, tree-based, and graph-based, which are accompanied by two categories of exploiting syntactic information: syntax pruning-based and syntax feature-based. Experiments are conducted on the CoNLL-2005, -2009, and -2012 benchmarks for all languages available, and results show that neural SRL models can still benefit from syntactic information under certain conditions. Furthermore, we show the quantitative significance of syntax to neural SRL models together with a thorough empirical survey using existing models.

QA-Align: Representing Cross-Text Content Overlap by Aligning Question-Answer Propositions

Daniela Brook Weiss et al. 11:45–12:00

Multi-text applications, such as multi-document summarization, are typically required to model redundancies across related texts. Current methods confronting consolidation struggle to fuse overlapping information. In order to explicitly represent content overlap, we propose to align predicate-argument relations across texts, providing a potential scaffold for information consolidation. We go beyond clustering coreferring mentions, and instead model overlap with respect to redundancy at a propositional level, rather than merely detecting shared referents. Our setting exploits QA-SRL, utilizing question-answer pairs to capture predicate-argument relations, facilitating laymen annotation of cross-text alignments. We employ crowd-workers for constructing a dataset of QA-based alignments, and present a baseline QA alignment model trained over our dataset. Analyses show that our new task is semantically challenging, capturing content overlap beyond lexical similarity and complements cross-document coreference with proposition-level links, offering potential use for downstream tasks.

PICARD: Parsing Incrementally for Constrained Auto-Regressive Decoding from Language Models
Torsten Scholak, Nathan Schucher, and Dzmitry Bahdanau 12:00–12:10

Large pre-trained language models for textual data have an unconstrained output space; at each decoding step, they can produce any of 10,000s of sub-word tokens. When fine-tuned to target constrained formal languages like SQL, these models often generate invalid code, rendering it unusable. We propose PICARD (code available at https://github.com/ElementAI/picard), a method for constraining auto-regressive decoders of language models through incremental parsing. PICARD helps to find valid output sequences by rejecting inadmissible tokens at each decoding step. On the challenging Spider and CoSQL text-to-SQL translation tasks, we show that PICARD transforms fine-tuned T5 models with passable performance into state-of-the-art solutions.

Exploiting Twitter as Source of Large Corpora of Weakly Similar Pairs for Semantic Sentence Embeddings
Marco Di Giovanni and Marco Brambilla 12:10–12:20

Semantic sentence embeddings are usually supervisedly built minimizing distances between pairs of embeddings of sentences labelled as semantically similar by annotators. Since big labelled datasets are rare, in particular for non-English languages, and expensive, recent studies focus on unsupervised approaches that require not-paired input sentences. We instead propose a language-independent approach to build large datasets of pairs of informal texts weakly similar, without manual human effort, exploiting Twitter's intrinsic powerful signals of relatedness: replies and quotes of tweets. We use the collected pairs to train a Transformer model with triplet-like structures, and we test the generated embeddings on Twitter NLP similarity tasks (PIT and TURL) and STSb. We also introduce four new sentence ranking evaluation benchmarks of informal texts, carefully extracted from the initial collections of tweets, proving not only that our best model learns classical Semantic Textual Similarity, but also excels on tasks where pairs of sentences are not exact paraphrases. Ablation studies reveal how increasing the corpus size influences positively the results, even at 2M samples, suggesting that bigger collections of Tweets still do not contain redundant information about semantic similarities. Code available at https://github.com/marco-digio/Twitter4SSE

Guilt by Association: Emotion Intensities in Lexical Representations
Shahab Raji and Gerard de Melo 12:20–12:30

What do linguistic models reveal about the emotions associated with words? In this study, we consider the task of estimating word-level emotion intensity scores for specific emotions, exploring unsupervised, supervised, and finally a self-supervised method of extracting emotional associations from pretrained vectors and models. Overall, we find that linguistic models carry substantial potential for inducing fine-grained emotion intensity scores, showing a far higher correlation with human ground truth ratings than state-of-the-art emotion lexicons based on labeled data.

Poster and Demo session 5

In-person Poster Session 5

Bavaro 4 *11:00–12:30*

Using Sociolinguistic Variables to Reveal Changing Attitudes Towards Sexuality and Gender

Sky CH-Wang and David Jurgens

Individuals signal aspects of their identity and beliefs through linguistic choices. Studying these choices in aggregate allows us to examine large-scale attitude shifts within a population. Here, we develop computational methods to study word choice within a sociolinguistic lexical variable—alternate words used to express the same concept—in order to test for change in the United States towards sexuality and gender. We examine two variables: i) referents to significant others, such as the word "partner" and ii) referents to an indefinite person, both of which could optionally be marked with gender. The linguistic choices in each variable allow us to study increased rates of acceptances of gay marriage and gender equality, respectively. In longitudinal analyses across Twitter and Reddit over 87M messages, we demonstrate that attitudes are changing but that these changes are driven by specific demographics within the United States. Further, in a quasi-causal analysis, we show that passages of Marriage Equality Acts in different states are drivers of linguistic change.

Identifying Morality Frames in Political Tweets using Relational Learning

Shamik Roy, Maria Leonor Pacheco, and Dan Goldwasser

Extracting moral sentiment from text is a vital component in understanding public opinion, social movements, and policy decisions. The Moral Foundation Theory identifies five moral foundations, each associated with a positive and negative polarity. However, moral sentiment is often motivated by its targets, which can correspond to individuals or collective entities. In this paper, we introduce morality frames, a representation framework for organizing moral attitudes directed at different entities, and come up with a novel and high-quality annotated dataset of tweets written by US politicians. Then, we propose a relational learning model to predict moral attitudes towards entities and moral foundations jointly. We do qualitative and quantitative evaluations, showing that moral sentiment towards entities differs highly across political ideologies.

Measuring Sentence-Level and Aspect-Level Certainty in Science Communications

Jiaxin Pei and David Jurgens

Certainty and uncertainty are fundamental to science communication. Hedges have widely been used as proxies for uncertainty. However, certainty is a complex construct, with authors expressing not only the degree but the type and aspects of uncertainty in order to give the reader a certain impression of what is known. Here, we introduce a new study of certainty that models both the level and the aspects of certainty in scientific findings. Using a new dataset of 2167 annotated scientific findings, we demonstrate that hedges alone account for only a partial explanation of certainty. We show that both the overall certainty and individual aspects can be predicted with pre-trained language models, providing a more complete picture of the author's intended communication. Downstream analyses on 431K scientific findings from news and scientific abstracts demonstrate that modeling sentence-level and aspect-level certainty is meaningful for areas like science communication. Both the model and datasets used in this paper are released at https://blablablab.si.umich.edu/projects/certainty/.

Assessing the Reliability of Word Embedding Gender Bias Measures

Yupei Du, Qixiang Fang, and Dong Nguyen

Various measures have been proposed to quantify human-like social biases in word embeddings. However, bias scores based on these measures can suffer from measurement error. One indication of measurement quality is reliability, concerning the extent to which a measure produces consistent results. In this paper, we assess three types of reliability of word embedding gender bias measures, namely test-retest reliability, inter-rater consistency and internal consistency. Specifically, we investigate the consistency of bias scores across different choices of random seeds, scoring rules and words. Furthermore, we analyse the effects of various factors on these measures' reliability scores. Our findings inform better design of word embedding gender bias measures. Moreover, we urge researchers to be more critical about the application of such measures

Rumor Detection on Twitter with Claim-Guided Hierarchical Graph Attention Networks

Hongzhan Lin et al.

Rumors are rampant in the era of social media. Conversation structures provide valuable clues to differentiate between real and fake claims. However, existing rumor detection methods are either limited to the strict relation of user responses or oversimplify the conversation structure. In this study, to substantially reinforces the interaction of user opinions while alleviating the negative impact imposed by irrelevant posts, we first represent the conversation thread as an undirected interaction graph. We then present a Claim-guided Hierarchical Graph Attention Network for rumor classification, which enhances the representation learning for responsive posts considering the entire social contexts and attends over the posts that can semantically infer the target claim. Extensive experiments on three Twitter datasets demonstrate that our rumor detection method achieves much better performance than state-of-the-art methods and exhibits a superior capacity for detecting rumors at early stages.

Learning Bill Similarity with Annotated and Augmented Corpora of Bills
Jiseon Kim et al.

Bill writing is a critical element of representative democracy. However, it is often overlooked that most legislative bills are derived, or even directly copied, from other bills. Despite the significance of bill-to-bill linkages for understanding the legislative process, existing approaches fail to address semantic similarities across bills, let alone reordering or paraphrasing which are prevalent in legal document writing. In this paper, we overcome these limitations by proposing a 5-class classification task that closely reflects the nature of the bill generation process. In doing so, we construct a human-labeled dataset of 4,721 bill-to-bill relationships at the subsection-level and release this annotated dataset to the research community. To augment the dataset, we generate synthetic data with varying degrees of similarity, mimicking the complex bill writing process. We use BERT variants and apply multi-stage training, sequentially fine-tuning our models with synthetic and human-labeled datasets. We find that the predictive performance significantly improves when training with both human-labeled and synthetic data. Finally, we apply our trained model to infer section- and bill-level similarities. Our analysis shows that the proposed methodology successfully captures the similarities across legal documents at various levels of aggregation.

SWEAT: Scoring Polarization of Topics across Different Corpora
Federico Bianchi et al.

Understanding differences of viewpoints across corpora is a fundamental task for computational social sciences. In this paper, we propose the Sliced Word Embedding Association Test (SWEAT), a novel statistical measure to compute the relative polarization of a topical wordset across two distributional representations. To this end, SWEAT uses two additional wordsets, deemed to have opposite valence, to represent two different poles. We validate our approach and illustrate a case study to show the usefulness of the introduced measure.

"So You Think You're Funny?": Rating the Humour Quotient in Standup Comedy
Anirudh Mittal et al.

Computational Humour (CH) has attracted the interest of Natural Language Processing and Computational Linguistics communities. Creating datasets for automatic measurement of humour quotient is difficult due to multiple possible interpretations of the content. In this work, we create a multi-modal humour-annotated dataset (~40 hours) using stand-up comedy clips. We devise a novel scoring mechanism to annotate the training data with a humour quotient score using the audience's laughter. The normalized duration (laughter duration divided by the clip duration) of laughter in each clip is used to compute this humour coefficient score on a five-point scale (0-4). This method of scoring is validated by comparing with manually annotated scores, wherein a quadratic weighted kappa of 0.6 is obtained. We use this dataset to train a model that provides a 'funniness' score, on a five-point scale, given the audio and its corresponding text. We compare various neural language models for the task of humour-rating and achieve an accuracy of 0.813 in terms of Quadratic Weighted Kappa (QWK). Our 'Open Mic' dataset is released for further research along with the code.

"Was it "stated" or was it "claimed"?: How linguistic bias affects generative language models
Roma Patel and Ellie Pavlick

People use language in subtle and nuanced ways to convey their beliefs. For instance, saying *claimed* instead of *said* casts doubt on the truthfulness of the underlying proposition, thus representing the author's opinion on the matter. Several works have identified such linguistic classes of words that occur frequently in natural language text and are bias-inducing by virtue of their framing effects. In this paper, we test whether generative language models (including GPT-2 [**radford2019language**] are sensitive to these linguistic framing effects. In particular, we test whether prompts that contain linguistic markers of author bias (e.g., hedges, implicatives, subjective intensifiers, assertives) influence the distribution of the generated text. Although these framing effects are subtle and stylistic, we find evidence that they lead to measurable style and topic differences in the generated text, leading to language that is, on average, more polarised and more skewed towards controversial entities and events.

PAUSE: Positive and Annealed Unlabeled Sentence Embedding
Lele Cao et al.

Sentence embedding refers to a set of effective and versatile techniques for converting raw text into numerical vector representations that can be used in a wide range of natural language processing (NLP) applications. The majority of these techniques are either supervised or unsupervised. Compared to the unsupervised methods, the supervised ones make less assumptions about optimization objectives and usually achieve better results. However, the training requires a large amount of labeled sentence pairs, which is not available in many industrial scenarios. To that end, we propose a generic and end-to-end approach – PAUSE (Positive and Annealed Unlabeled Sentence Embedding), capable of learning high-quality sentence embeddings from a partially labeled dataset. We experimentally show that PAUSE achieves, and sometimes surpasses, state-of-the-art results using only a small fraction of labeled sentence pairs on various benchmark tasks. When applied to a real industrial use case where labeled samples are scarce, PAUSE encourages us to extend our dataset without the burden of extensive manual annotation work.

A Simple Geometric Method for Cross-Lingual Linguistic Transformations with Pre-trained Autoencoders
Maarten De Raedt et al.

Powerful sentence encoders trained for multiple languages are on the rise. These systems are capable of embedding a wide range of linguistic properties into vector representations. While explicit probing tasks can be used to verify the presence of specific linguistic properties, it is unclear whether the vector representations can be manipulated to indirectly steer such properties. For efficient learning, we investigate the use of a geometric mapping in embedding space to transform linguistic properties, without any tuning of the pre-trained sentence encoder or decoder. We validate our approach on three linguistic properties using a pre-trained multilingual autoencoder and analyze the results in both monolingual and cross-lingual settings.

An Information-Theoretic Characterization of Morphological Fusion
Neil Rathi, Michael Hahn, and Richard Futrell

Linguistic typology generally divides synthetic languages into groups based on their morphological fusion. However, this measure has long been thought to be best considered a matter of degree. We present an information-theoretic measure, called informational fusion, to quantify the degree of fusion of a given set of morphological features in a surface form, which naturally provides such a graded scale. Informational fusion is able to encapsulate not only concatenative, but also nonconcatenative morphological systems (e.g. Arabic), abstracting away from any notions of morpheme segmentation. We then show, on a sample of twenty-one languages, that our measure recapitulates the usual linguistic classifications for concatenative systems, and provides new measures for nonconcatenative ones. We also evaluate the long-standing hypotheses that more frequent forms are more fusional, and that paradigm size anticorrelates with degree of fusion. We do not find evidence for the idea that languages have characteristic levels of fusion; rather, the degree of fusion varies across part-of-speech within languages.

The Effect of Efficient Messaging and Input Variability on Neural-Agent Iterated Language Learning
Yuchen Lian, Arianna Bisazza, and Tessa Verhoef

Natural languages display a trade-off among different strategies to convey syntactic structure, such as word order or inflection. This trade-off, however, has not appeared in recent simulations of iterated language learning with neural network agents (Chaabouni et al., 2019b). We re-evaluate this result in light of three factors that play an important role in comparable experiments from the Language Evolution field: (i) speaker bias towards efficient messaging, (ii) non systematic input languages, and (iii) learning bottleneck. Our simulations show that neural agents mainly strive to maintain the utterance type distribution observed during learning, instead of developing a more efficient or systematic language.

On Classifying whether Two Texts are on the Same Side of an Argument
Erik Körner et al.

To ease the difficulty of argument stance classification, the task of same side stance classification (S3C) has been proposed. In contrast to actual stance classification, which requires a substantial amount of domain knowledge to identify whether an argument is in favor or against a certain issue, it is argued that, for S3C, only argument similarity within stances needs to be learned to successfully solve the task. We evaluate several transformer-based approaches on the dataset of the recent S3C shared task, followed by an in-depth evaluation and error analysis of our model and the task's hypothesis. We show that, although we achieve state-of-the-art results, our model fails to generalize both within as well as across topics and domains when adjusting the sampling strategy of the training and test set to a more adversarial scenario. Our evaluation shows that current state-of-the-art approaches cannot determine same side stance by considering only domain-independent linguistic similarity features, but appear to require domain knowledge and semantic inference, too.

Chinese Opinion Role Labeling with Corpus Translation: A Pivot Study
Ranran Zhen et al.

Opinion Role Labeling (ORL), aiming to identify the key roles of opinion, has received increasing interest. Unlike most of the previous works focusing on the English language, in this paper, we present the first work of Chinese ORL. We construct a Chinese dataset by manually translating and projecting annotations from a standard English MPQA dataset. Then, we investigate the effectiveness of cross-lingual transfer methods, including model transfer and corpus translation. We exploit multilingual BERT with Contextual Parameter Generator and Adapter methods to examine the potentials of unsupervised cross-lingual learning and our experiments and analyses for both bilingual and multilingual transfers establish a foundation for the future research of this task.

MassiveSumm: a very large-scale, very multilingual, news summarisation dataset
Daniel Varab and Natalie Schluter

Current research in automatic summarisation is unapologetically anglo-centered–a persistent state-of-affairs, which also predates neural net approaches. High-quality automatic summarisation datasets are notoriously expensive to create, posing a challenge for any language. However, with digitalisation, archiving, and social media advertising of newswire articles, recent work has shown how, with careful methodology application, large-scale datasets can now be simply gathered instead of written. In this paper, we present a large-scale multilingual summarisation dataset containing articles in 92 languages, spread across 28.8 million articles, in more than 35 writing scripts. This is both the largest, most inclusive, existing automatic summarisation dataset, as well as one of the largest, most inclusive, ever published datasets for any NLP task. We present the first investigation on the efficacy of resource building from news platforms in the low-resource language setting. Finally, we provide some first insight on how low-resource language settings impact state-of-the-art automatic

summarisation system performance.

AUTOSUMM: Automatic Model Creation for Text Summarization
Sharmila Reddy Nangi et al.

Recent efforts to develop deep learning models for text generation tasks such as extractive and abstractive summarization have resulted in state-of-the-art performances on various datasets. However, obtaining the best model configuration for a given dataset requires an extensive knowledge of deep learning specifics like model architecture, tuning parameters etc., and is often extremely challenging for a non-expert. In this paper, we propose methods to automatically create deep learning models for the tasks of extractive and abstractive text summarization. Based on the recent advances in Automated Machine Learning and the success of large language models such as BERT and GPT-2 in encoding knowledge, we use a combination of Neural Architecture Search (NAS) and Knowledge Distillation (KD) techniques to perform model search and compression using the vast knowledge provided by these language models to develop smaller, customized models for any given dataset. We present extensive empirical results to illustrate the effectiveness of our model creation methods in terms of inference time and model size, while achieving near state-of-the-art performances in terms of accuracy across a range of datasets.

Idiosyncratic but not Arbitrary: Learning Idiolects in Online Registers Reveals Distinctive yet Consistent Individual Styles
Jian Zhu and David Jurgens

An individual's variation in writing style is often a function of both social and personal attributes. While structured social variation has been extensively studied, e.g., gender based variation, far less is known about how to characterize individual styles due to their idiosyncratic nature. We introduce a new approach to studying idiolects through a massive cross-author comparison to identify and encode stylistic features. The neural model achieves strong performance at authorship identification on short texts and through an analogy-based probing task, showing that the learned representations exhibit surprising regularities that encode qualitative and quantitative shifts of idiolectal styles. Through text perturbation, we quantify the relative contributions of different linguistic elements to idiolectal variation. Furthermore, we provide a description of idiolects through measuring inter- and intra-author variation, showing that variation in idiolects is often distinctive yet consistent.

Efficient Multi-Task Auxiliary Learning: Selecting Auxiliary Data by Feature Similarity
Po-Nien Kung et al.

Multi-task auxiliary learning utilizes a set of relevant auxiliary tasks to improve the performance of a primary task. A common usage is to manually select multiple auxiliary tasks for multi-task learning on all data, which raises two issues: (1) selecting beneficial auxiliary tasks for a primary task is nontrivial; (2) when the auxiliary datasets are large, training on all data becomes time-expensive and impractical. Therefore, this paper focuses on addressing these problems and proposes a time-efficient sampling method to select the data that is most relevant to the primary task. The proposed method allows us to only train on the most beneficial sub-datasets from the auxiliary tasks, achieving efficient multi-task auxiliary learning. The experiments on three benchmark datasets (RTE, MRPC, STS-B) show that our method significantly outperforms random sampling and ST-DNN. Also, by applying our method, the model can surpass fully-trained MT-DNN on RTE, MRPC, STS-B, using only 50%, 66%, and 1% of data, respectively.

Distilling Linguistic Context for Language Model Compression
Geondo Park, Gyeongman Kim, and Eunho Yang

A computationally expensive and memory intensive neural network lies behind the recent success of language representation learning. Knowledge distillation, a major technique for deploying such a vast language model in resource-scarce environments, transfers the knowledge on individual word representations learned without restrictions. In this paper, inspired by the recent observations that language representations are relatively positioned and have more semantic knowledge as a whole, we present a new knowledge distillation objective for language representation learning that transfers the contextual knowledge via two types of relationships across representations: Word Relation and Layer Transforming Relation. Unlike other recent distillation techniques for the language models, our contextual distillation does not have any restrictions on architectural changes between teacher and student. We validate the effectiveness of our method on challenging benchmarks of language understanding tasks, not only in architectures of various sizes but also in combination with DynaBERT, the recently proposed adaptive size pruning method.

Beyond Preserved Accuracy: Evaluating Loyalty and Robustness of BERT Compression
Canwen Xu et al.

Recent studies on compression of pretrained language models (e.g., BERT) usually use preserved accuracy as the metric for evaluation. In this paper, we propose two new metrics, label loyalty and probability loyalty that measure how closely a compressed model (i.e., student) mimics the original model (i.e., teacher). We also explore the effect of compression with regard to robustness under adversarial attacks. We benchmark quantization, pruning, knowledge distillation and progressive module replacing with loyalty and robustness. By combining multiple compression techniques, we provide a practical strategy to achieve better accuracy, loyalty and robustness.

A Secure and Efficient Federated Learning Framework for NLP
Chenghong Wang et al.

In this work, we consider the problem of designing secure and efficient federated learning (FL) frameworks for NLP. Existing solutions under this literature either consider a trusted aggregator or require heavy-weight cryptographic primitives, which makes the performance significantly degraded. Moreover, many existing secure FL designs work only under the restrictive assumption that none of the clients can be dropped out from the training protocol. To tackle these problems, we propose SEFL, a secure and efficient federated learning framework that (1)~eliminates the need for the trusted entities; (2)~achieves similar and even better model accuracy compared with existing FL designs; (3)~is resilient to client dropouts.

On Homophony and Rényi Entropy
Tiago Pimentel et al.

Homophony's widespread presence in natural languages is a controversial topic. Recent theories of language optimality have tried to justify its prevalence, despite its negative effects on cognitive processing time, e.g., Piantadosi et al. (2012) argued homophony enables the reuse of efficient wordforms and is thus beneficial for languages. This hypothesis has recently been challenged by Trott and Bergen (2020), who posit that good wordforms are more often homophonous simply because they are more phonotactically probable. In this paper, we join in on the debate. We first propose a new information-theoretic quantification of a language's homophony: the sample Rényi entropy. Then, we use this quantification to revisit Trott and Bergen's claims. While their point is theoretically sound, a specific methodological issue in their experiments raises doubts about their results. After addressing this issue, we find no clear pressure either towards or against homophony—a much more nuanced result than either Piantadosi et al.'s or Trott and Bergen's findings.

Few-Shot Emotion Recognition in Conversation with Sequential Prototypical Networks
Gaël Guibon et al.

Several recent studies on dyadic human-human interactions have been done on conversations without specific business objectives. However, many companies might benefit from studies dedicated to more precise environments such as after sales services or customer satisfaction surveys. In this work, we place ourselves in the scope of a live chat customer service in which we want to detect emotions and their evolution in the conversation flow. This context leads to multiple challenges that range from exploiting restricted, small and mostly unlabeled datasets to finding and adapting methods for such context. We tackle these challenges by using Few-Shot Learning while making the hypothesis it can serve conversational emotion classification for different languages and sparse labels. We contribute by proposing a variation of Prototypical Networks for sequence labeling in conversation that we name ProtoSeq. We test this method on two datasets with different languages: daily conversations in English and customer service chat conversations in French. When applied to emotion classification in conversations, our method proved to be competitive even when compared to other ones.

Chinese Opinion Role Labeling with Corpus Translation: A Pivot Study
Ranran Zhen et al.

Opinion Role Labeling (ORL), aiming to identify the key roles of opinion, has received increasing interest. Unlike most of the previous works focusing on the English language, in this paper, we present the first work of Chinese ORL. We construct a Chinese dataset by manually translating and projecting annotations from a standard English MPQA dataset. Then, we investigate the effectiveness of cross-lingual transfer methods, including model transfer and corpus translation. We exploit multilingual BERT with Contextual Parameter Generator and Adapter methods to examine the potentials of unsupervised cross-lingual learning and our experiments and analyses for both bilingual and multilingual transfers establish a foundation for the future research of this task.

Unsupervised Abstractive Opinion Summarization by Generating Sentences with Tree-Structured Topic Guidance
Masaru Isonuma et al.

This paper presents a novel unsupervised abstractive summarization method for opinionated texts. While the basic variational autoencoder-based models assume a unimodal Gaussian prior for the latent code of sentences, we alternate it with a recursive Gaussian mixture, where each mixture component corresponds to the latent code of a topic sentence and is mixed by a tree-structured topic distribution. By decoding each Gaussian component, we generate sentences with tree-structured topic guidance, where the root sentence conveys generic content, and the leaf sentences describe specific topics. Experimental results demonstrate that the generated topic sentences are appropriate as a summary of opinionated texts, which are more informative and cover more input contents than those generated by the recent unsupervised summarization model (Brazinskas et al., 2020). Furthermore, we demonstrate that the variance of latent Gaussians represents the granularity of sentences, analogous to Gaussian word embedding (Vilnis and McCallum, 2015).

In-person Demo Session

Bavaro 4 *11:00–12:30*

Press Freedom Monitor: Detection of Reported Press and Media Freedom Violations in Twitter and News Articles
Tariq Yousef et al.

Freedom of the press and media is of vital importance for democratically organised states and open societies. We introduce the Press Freedom Monitor, a tool that aims to detect reported press and media freedom violations in news articles and tweets. It is used by press and media freedom organisations to support their daily monitoring and to trigger rapid response actions. The Press Freedom Monitor enables the monitoring experts to get a fast overview over recently reported incidents and it has shown an impressive performance in this regard. This paper presents our work on the tool, starting with the training phase, which comprises defining the topic-related keywords to be used for querying APIs for news and Twitter content and evaluating different machine learning models based on a training dataset specifically created for our use case. Then, we describe the components of the production pipeline, including data gathering, duplicates removal, country mapping, case mapping and the user interface. We also conducted a usability study to evaluate the effectiveness of the user interface, and describe improvement plans for future work.

Project Debater APIs: Decomposing the AI Grand Challenge
Roy Bar-Haim et al.

Project Debater was revealed in 2019 as the first AI system that can debate human experts on complex topics. Engaging in a live debate requires a diverse set of skills, and Project Debater has been developed accordingly as a collection of components, each designed to perform a specific subtask. Project Debater APIs provide access to many of these capabilities, as well as to more recently developed ones. This diverse set of web services, publicly available for academic use, includes core NLP services, argument mining and analysis capabilities, and higher-level services for content summarization. We describe these APIs and their performance, and demonstrate how they can be used for building practical solutions. In particular, we will focus on Key Point Analysis, a novel technology that identifies the main points and their prevalence in a collection of texts such as survey responses and user reviews.

InVeRo-XL: Making Cross-Lingual Semantic Role Labeling Accessible with Intelligible Verbs and Roles
Simone Conia et al.

Notwithstanding the growing interest in cross-lingual techniques for Natural Language Processing, there has been a surprisingly small number of efforts aimed at the development of easy-to-use tools for cross-lingual Semantic Role Labeling. In this paper, we fill this gap and present InVeRo-XL, an off-the-shelf state-of-the-art system capable of annotating text with predicate sense and semantic role labels from 7 predicate-argument structure inventories in more than 40 languages. We hope that our system, with its easy-to-use RESTful API and Web interface, will become a valuable tool for the research community, encouraging the integration of sentence-level semantics into cross-lingual downstream tasks. InVeRo-XL is available online at http://nlp.uniroma1.it/invero.

Session 9 Overview – Tuesday, November 9, 2021

Track A	Track B	Track C	Track D	Track E	Track F	
Machine Translation and Multilinguality 4	*Interpretability and Analysis of Models for NLP 3*	*Information Extraction 5*	*Resources and Evaluation 4*	*Syntax*	*Efficient Methods for NLP 3*	
Investigating the Helpfulness of Word-Level Quality Estimation for Post-Editing Machine Translation Output *Shenoy et al.*	Measuring Association Between Labels and Free-Text Rationales *Wiegreffe, Marasović, and Smith*	Distantly-Supervised Named Entity Recognition with Noise-Robust Learning and Language Model Augmented Self-Training *Meng et al.*	Automatic Text Evaluation through the Lens of Wasserstein Barycenters *Colombo et al.*	A Root of a Problem: Optimizing Single-Root Dependency Parsing *Stanojević and Cohen*	What to Pre-Train on? Efficient Intermediate Task Selection *Poth et al.*	2:30
UNKs Everywhere: Adapting Multilingual Language Models to New Scripts *Pfeiffer et al.*	Discretized Integrated Gradients for Explaining Language Models *Sanyal and Ren*	Open Knowledge Graphs Canonicalization using Variational Autoencoders *Dash et al.*	Visually Grounded Reasoning across Languages and Cultures *Liu et al.*	Efficient Sampling of Dependency Structure *Zmigrod, Vieira, and Cotterell*	PermuteFormer: Efficient Relative Position Encoding for Long Sequences *Chen*	2:45
Neural Machine Translation Quality and Post-Editing Performance *Zouhar et al.*	Putting Words in BERT's Mouth: Navigating Contextualized Vector Spaces with Pseudowords *Karidi et al.*	HittER: Hierarchical Transformers for Knowledge Graph Embeddings *Chen et al.*	Back to Square One: Artifact Detection, Training and Commonsense Disentanglement in the Winograd Schema *Elazar et al.*	Efficient Computation of Expectations under Spanning Tree Distributions *Zmigrod, Vieira, and Cotterell*	Block Pruning For Faster Transformers *Lagunas et al.*	3:00
XTREME-R: Towards More Challenging and Nuanced Multilingual Evaluation *Ruder et al.*	Rationales for Sequential Predictions *Vafa et al.*	Few-Shot Named Entity Recognition: An Empirical Baseline Study *Huang et al.*	Robustness Evaluation of Entity Disambiguation Using Prior Probes: the Case of Entity Overshadowing *Provatorova et al.*	Joint Universal Syntactic and Semantic Parsing *Stengel-Eskin et al.*	Finetuning Pretrained Transformers into RNNs *Kasai et al.*	3:15

	Track A	**Track B**	**Track C**	**Track D**	**Track E**	**Track F**
	Machine Translation and Multilinguality 4	*Interpretability and Analysis of Models for NLP 3*	*Information Extraction 5*	*Resources and Evaluation 4*	*Syntax*	*Efficient Methods for NLP 3*
3:30	Contrastive Conditioning for Assessing Disambiguation in MT: A Case Study of Distilled Bias *Vamvas and Sennrich*	FastIF: Scalable Influence Functions for Efficient Model Interpretation and Debugging *Guo et al.*	XLEnt: Mining a Large Cross-lingual Entity Dataset with Lexical-Semantic-Phonetic Word Alignment *El-Kishky et al.*	IndoNLI: A Natural Language Inference Dataset for Indonesian *Mahendra et al.*	Reducing Discontinuous to Continuous Parsing with Pointer Network Reordering *Fernández-González and Gómez-Rodríguez*	How to Train BERT with an Academic Budget *Izsak, Berchansky, and Levy*
3:45	On the Difficulty of Translating Free-Order Case-Marking Languages *Bisazza, Üstün, and Sportel*	Studying word order through iterative shuffling *Malkin et al.*	Utilizing Relative Event Time to Enhance Event-Event Temporal Relation Extraction *Wen and Ji*	Agreeing to Disagree: Annotating Offensive Language Datasets with Annotators' Disagreement *Leonardelli et al.*	A New Representation for Span-based CCG Parsing *Kato and Matsubara*	Beyond Preserved Accuracy: Evaluating Loyalty and Robustness of BERT Compression *Xu et al.*
3:50			Separating Retention from Extraction in the Evaluation of End-to-end Relation Extraction *Taillé et al.*			IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization *Koto, Lau, and Baldwin*

Parallel Session 9

Session 9A: Machine Translation and Multilinguality 4

Bavaro 1 *Chair: Sashi Narayan*

Investigating the Helpfulness of Word-Level Quality Estimation for Post-Editing Machine Translation Output

Raksha Shenoy et al. 14:30–14:45

Compared to fully manual translation, post-editing (PE) machine translation (MT) output can save time and reduce errors. Automatic word-level quality estimation (QE) aims to predict the correctness of words in MT output and holds great promise to aid PE by flagging problematic output. Quality of QE is crucial, as incorrect QE might lead to translators missing errors or wasting time on already correct MT output. Achieving accurate automatic word-level QE is very hard, and it is currently not known (i) at what quality threshold QE is actually beginning to be useful for human PE, and (ii), how to best present word-level QE information to translators. In particular, should word-level QE visualization indicate uncertainty of the QE model or not? In this paper, we address both research questions with real and simulated word-level QE, visualizations, and user studies, where time, subjective ratings, and quality of the final translations are assessed. Results show that current word-level QE models are not yet good enough to support PE. Instead, quality levels of > 80% F1 are required. For helpful quality levels, a visualization reflecting the uncertainty of the QE model is preferred. Our analysis further shows that speed gains achieved through QE are not merely a result of blindly trusting the QE system, but that the quality of the final translations also improves. The threshold results from the paper establish a quality goal for future word-level QE research.

UNKs Everywhere: Adapting Multilingual Language Models to New Scripts

Jonas Pfeiffer et al. 14:45–15:00

Massively multilingual language models such as multilingual BERT offer state-of-the-art cross-lingual transfer performance on a range of NLP tasks. However, due to limited capacity and large differences in pretraining data sizes, there is a profound performance gap between resource-rich and resource-poor target languages. The ultimate challenge is dealing with under-resourced languages not covered at all by the models and written in scripts unseen during pretraining. In this work, we propose a series of novel data-efficient methods that enable quick and effective adaptation of pretrained multilingual models to such low-resource languages and unseen scripts. Relying on matrix factorization, our methods capitalize on the existing latent knowledge about multiple languages already available in the pretrained model's embedding matrix. Furthermore, we show that learning of the new dedicated embedding matrix in the target language can be improved by leveraging a small number of vocabulary items (i.e., the so-called lexically overlapping tokens) shared between mBERT's and target language vocabulary. Our adaptation techniques offer substantial performance gains for languages with unseen scripts. We also demonstrate that they can yield improvements for low-resource languages written in scripts covered by the pretrained model.

Neural Machine Translation Quality and Post-Editing Performance

Vilém Zouhar et al. 15:00–15:15

We test the natural expectation that using MT in professional translation saves human processing time. The last such study was carried out by Sanchez-Torron and Koehn (2016) with phrase-based MT, artificially reducing the translation quality. In contrast, we focus on neural MT (NMT) of high quality, which has become the state-of-the-art approach since then and also got adopted by most translation companies. Through an experimental study involving over 30 professional translators for English -> Czech translation, we examine the relationship between NMT performance and post-editing time and quality. Across all models, we found that better MT systems indeed lead to fewer changes in the sentences in this industry setting. The relation between system quality and post-editing time is however not straightforward and, contrary to the results on phrase-based MT, BLEU is definitely not a stable predictor of the time or final output quality.

XTREME-R: Towards More Challenging and Nuanced Multilingual Evaluation

Sebastian Ruder et al. 15:15–15:30

Machine learning has brought striking advances in multilingual natural language processing capabilities over the past year. For example, the latest techniques have improved the state-of-the-art performance on the XTREME multilingual benchmark by more than 13 points. While a sizeable gap to human-level performance remains, improvements have been easier to achieve in some tasks than in others. This paper analyzes the current state of cross-lingual transfer learning and summarizes some lessons learned. In order to catalyze meaningful progress, we extend XTREME to XTREME-R, which consists of an improved set of ten natural language understanding tasks, including challenging language-agnostic retrieval tasks, and covers 50 typologically diverse languages. In addition, we provide a massively multilingual diagnostic suite and fine-grained multi-dataset evaluation capabilities through an interactive public leaderboard to gain a better understanding of such models.

Contrastive Conditioning for Assessing Disambiguation in MT: A Case Study of Distilled Bias

Jannis Vamvas and Rico Sennrich 15:30–15:45

Lexical disambiguation is a major challenge for machine translation systems, especially if some senses of a word are trained less often than others. Identifying patterns of overgeneralization requires evaluation methods that are both reliable and scalable. We propose contrastive conditioning as a reference-free black-box method for detecting disambiguation errors. Specifically, we score the quality of a translation by conditioning on variants of the source that provide contrastive disambiguation cues. After validating our method, we apply it in a case study to perform a targeted evaluation of sequence-level knowledge distillation. By probing word sense disambiguation and translation of gendered occupation names, we show that distillation-trained models tend to overgeneralize more than other models with a comparable BLEU score. Contrastive conditioning thus highlights a side effect of distillation that is not fully captured by standard evaluation metrics. Code and data to reproduce our findings are publicly available.

On the Difficulty of Translating Free-Order Case-Marking Languages

Arianna Bisazza, Ahmet Üstün, and Stephan Sportel 15:45–16:00

Identifying factors that make certain languages harder to model than others is essential to reach language equality in future Natural Language Processing technologies. Free-order case-marking languages, such as Russian, Latin or Tamil, have proved more challenging than fixed-order languages for the tasks of syntactic parsing and subject-verb agreement prediction. In this work, we investigate whether this class of languages is also more difficult to translate by state-of-the-art Neural Machine Translation models (NMT). Using a variety of synthetic languages and a newly introduced translation challenge set, we find that word order flexibility in the source language only leads to a very small loss of NMT quality, even though the core verb arguments become impossible to disambiguate in sentences without semantic cues. The latter issue is indeed solved by the addition of case marking. However, in medium- and low-resource settings, the overall NMT quality of fixed-order languages remains unmatched.

Session 9B: Interpretability and Analysis of Models for NLP 3

Bavaro 2 *Chair: Yonatan Belinkov*

Measuring Association Between Labels and Free-Text Rationales

Sarah Wiegreffe, Ana Marasović, and Noah A. Smith 14:30–14:45

In interpretable NLP, we require faithful rationales that reflect the model's decision-making process for an explained instance. While prior work focuses on extractive rationales (a subset of the input words), we investigate their less-studied counterpart: free-text natural language rationales. We demonstrate that *pipelines*, models for faithful rationalization on information-extraction style tasks, do not work as well on "reasoning" tasks requiring free-text rationales. We turn to models that *jointly* predict and rationalize, a class of widely used high-performance models for free-text rationalization. We investigate the extent to which the labels and rationales predicted by these models are associated, a necessary property of faithful explanation. Via two tests, *robustness equivalence* and *feature importance agreement*, we find that state-of-the-art T5-based joint models exhibit desirable properties for explaining commonsense question-answering and natural language inference, indicating their potential for producing faithful free-text rationales.

Discretized Integrated Gradients for Explaining Language Models

Soumya Sanyal and Xiang Ren 14:45–15:00

As a prominent attribution-based explanation algorithm, Integrated Gradients (IG) is widely adopted due to its desirable explanation axioms and the ease of gradient computation. It measures feature importance by averaging the model's output gradient interpolated along a straight-line path in the input data space. However, such straight-line interpolated points are not representative of text data due to the inherent discreteness of the word embedding space. This questions the faithfulness of the gradients computed at the interpolated points and consequently, the quality of the generated explanations. Here we propose Discretized Integrated Gradients (DIG), which allows effective attribution along non-linear interpolation paths. We develop two interpolation strategies for the discrete word embedding space that generates interpolation points that lie close to actual words in the embedding space, yielding more faithful gradient computation. We demonstrate the effectiveness of DIG over IG through experimental and human evaluations on multiple sentiment classification datasets. We provide the source code of DIG to encourage reproducible research.

Putting Words in BERT's Mouth: Navigating Contextualized Vector Spaces with Pseudowords

Taelin Karidi et al. 15:00–15:15

We present a method for exploring regions around individual points in a contextualized vector space (particularly, BERT space), as a way to investigate how these regions correspond to word senses. By inducing a contextualized "pseudoword" vector as a stand-in for a static embedding in the input layer, and then performing masked prediction of a word in the sentence, we are able to investigate the geometry of the BERT-space in a controlled manner around individual instances. Using our method on a set of carefully constructed sentences targeting highly ambiguous English words, we find substantial regularity in the contextualized space, with regions that correspond to distinct word senses; but between these regions there are occasionally "sense voids"—regions that do not correspond to any intelligible sense.

Rationales for Sequential Predictions

Keyon Vafa et al. 15:15–15:30

Sequence models are a critical component of modern NLP systems, but their predictions are difficult to explain. We consider model explanations though rationales, subsets of context that can explain individual model predictions. We find sequential rationales by solving a combinatorial optimization: the best rationale is the smallest subset of input tokens that would predict the same output as the full sequence. Enumerating all subsets is intractable, so we propose an efficient greedy algorithm to approximate this objective. The algorithm, which is called greedy rationalization, applies to any model. For this approach to be effective, the model should form compatible conditional distributions when making predictions on incomplete subsets of the context. This condition can be enforced with a short fine-tuning step. We study greedy rationalization on language modeling and machine translation. Compared to existing baselines, greedy rationalization is best at optimizing the sequential objective and provides the most faithful rationales. On a new dataset of annotated sequential rationales, greedy rationales are most similar to human rationales.

FastIF: Scalable Influence Functions for Efficient Model Interpretation and Debugging

Han Guo et al. 15:30–15:45

Influence functions approximate the "influences" of training data-points for test predictions and have a wide variety of applications. Despite the popularity, their computational cost does not scale well with model and training data size. We present FastIF, a set of simple modifications to influence functions that significantly improves their run-time. We use k-Nearest Neighbors (kNN) to narrow the search space down to a subset of good candidate data points, identify the configurations that best balance the speed-quality trade-off in estimating the inverse Hessian-vector product, and introduce a fast parallel variant. Our proposed method achieves about 80X speedup while being highly correlated with the original influence values. With the availability of the fast influence functions, we demonstrate their usefulness in four applications. First, we examine whether influential data-points can "explain" test time behavior using the framework of simulatability. Second, we visualize the influence interactions between training and test data-points. Third, we show that we can correct model errors by additional fine-tuning on certain influential data-points, improving the accuracy of a trained MultiNLI

model by 2.5% on the HANS dataset. Finally, we experiment with a similar setup but fine-tuning on datapoints not seen during training, improving the model accuracy by 2.8% and 1.7% on HANS and ANLI datasets respectively. Overall, our fast influence functions can be efficiently applied to large models and datasets, and our experiments demonstrate the potential of influence functions in model interpretation and correcting model errors.

Studying word order through iterative shuffling
Nikolay Malkin et al. 15:45–16:00

As neural language models approach human performance on NLP benchmark tasks, their advances are widely seen as evidence of an increasingly complex understanding of syntax. This view rests upon a hypothesis that has not yet been empirically tested: that word order encodes meaning essential to performing these tasks. We refute this hypothesis in many cases: in the GLUE suite and in various genres of English text, the words in a sentence or phrase can rarely be permuted to form a phrase carrying substantially different information. Our surprising result relies on inference by iterative shuffling (IBIS), a novel, efficient procedure that finds the ordering of a bag of words having the highest likelihood under a fixed language model. IBIS can use any black-box model without additional training and is superior to existing word ordering algorithms. Coalescing our findings, we discuss how shuffling inference procedures such as IBIS can benefit language modeling and constrained generation.

Session 9C: Information Extraction 5

Bavaro 3 *Chair: Yulan He*

Distantly-Supervised Named Entity Recognition with Noise-Robust Learning and Language Model Augmented Self-Training

Yu Meng et al. 14:30–14:45

We study the problem of training named entity recognition (NER) models using only distantly-labeled data, which can be automatically obtained by matching entity mentions in the raw text with entity types in a knowledge base. The biggest challenge of distantly-supervised NER is that the distant supervision may induce incomplete and noisy labels, rendering the straightforward application of supervised learning ineffective. In this paper, we propose (1) a noise-robust learning scheme comprised of a new loss function and a noisy label removal step, for training NER models on distantly-labeled data, and (2) a self-training method that uses contextualized augmentations created by pre-trained language models to improve the generalization ability of the NER model. On three benchmark datasets, our method achieves superior performance, outperforming existing distantly-supervised NER models by significant margins.

Open Knowledge Graphs Canonicalization using Variational Autoencoders

Sarthak Dash et al. 14:45–15:00

Noun phrases and Relation phrases in open knowledge graphs are not canonicalized, leading to an explosion of redundant and ambiguous subject-relation-object triples. Existing approaches to solve this problem take a two-step approach. First, they generate embedding representations for both noun and relation phrases, then a clustering algorithm is used to group them using the embeddings as features. In this work, we propose Canonicalizing Using Variational AutoEncoders and Side Information (CUVA), a joint model to learn both embeddings and cluster assignments in an end-to-end approach, which leads to a better vector representation for the noun and relation phrases. Our evaluation over multiple benchmarks shows that CUVA outperforms the existing state-of-the-art approaches. Moreover, we introduce CanonicNell, a novel dataset to evaluate entity canonicalization systems.

HittER: Hierarchical Transformers for Knowledge Graph Embeddings

Sanxing Chen et al. 15:00–15:15

This paper examines the challenging problem of learning representations of entities and relations in a complex multi-relational knowledge graph. We propose HittER, a Hierarchical Transformer model to jointly learn Entity-relation composition and Relational contextualization based on a source entity's neighborhood. Our proposed model consists of two different Transformer blocks: the bottom block extracts features of each entity-relation pair in the local neighborhood of the source entity and the top block aggregates the relational information from outputs of the bottom block. We further design a masked entity prediction task to balance information from the relational context and the source entity itself. Experimental results show that HittER achieves new state-of-the-art results on multiple link prediction datasets. We additionally propose a simple approach to integrate HittER into BERT and demonstrate its effectiveness on two Freebase factoid question answering datasets.

Few-Shot Named Entity Recognition: An Empirical Baseline Study

Jiaxin Huang et al. 15:15–15:30

This paper presents an empirical study to efficiently build named entity recognition (NER) systems when a small amount of in-domain labeled data is available. Based upon recent Transformer-based self-supervised pre-trained language models (PLMs), we investigate three orthogonal schemes to improve model generalization ability in few-shot settings: (1) meta-learning to construct prototypes for different entity types, (2) task-specific supervised pre-training on noisy web data to extract entity-related representations and (3) self-training to leverage unlabeled in-domain data. On 10 public NER datasets, we perform extensive empirical comparisons over the proposed schemes and their combinations with various proportions of labeled data, our experiments show that (i)in the few-shot learning setting, the proposed NER schemes significantly improve or outperform the commonly used baseline, a PLM-based linear classifier fine-tuned using domain labels. (ii) We create new state-of-the-art results on both few-shot and training-free settings compared with existing methods.

XLEnt: Mining a Large Cross-lingual Entity Dataset with Lexical-Semantic-Phonetic Word Alignment

Ahmed El-Kishky et al. 15:30–15:40

Cross-lingual named-entity lexica are an important resource to multilingual NLP tasks such as machine translation and cross-lingual wikification. While knowledge bases contain a large number of entities in high-resource languages such as English and French, corresponding entities for lower-resource languages are often missing. To address this, we propose Lexical-Semantic-Phonetic Align (LSP-Align), a technique to automatically mine cross-lingual entity lexica from mined web data. We demonstrate LSP-Align outperforms baselines at extracting cross-lingual entity pairs and mine 164 million entity pairs from 120 different languages aligned with English. We release these cross-lingual entity pairs along with the massively multilingual tagged named entity corpus as a resource to the NLP community.

Utilizing Relative Event Time to Enhance Event-Event Temporal Relation Extraction

Haoyang Wen and Heng Ji 15:40–15:50

Event time is one of the most important features for event-event temporal relation extraction. However, explicit event time information in text is sparse. For example, only about 20% of event mentions in TimeBank-Dense have event-time links. In this paper, we propose a joint model for event-event temporal relation classification and an auxiliary task, relative event time prediction, which predicts the event time as real numbers. We adopt the Stack-Propagation framework to incorporate predicted relative event time for temporal relation classification and keep the differentiability. Our experiments on MATRES dataset show that our model can significantly improve the RoBERTa-based baseline and achieve state-of-the-art performance.

Separating Retention from Extraction in the Evaluation of End-to-end Relation Extraction
Bruno Taillé et al. 15:50–16:00

State-of-the-art NLP models can adopt shallow heuristics that limit their generalization capability (McCoy et al., 2019). Such heuristics include lexical overlap with the training set in Named-Entity Recognition (Taille et al., 2020) and Event or Type heuristics in Relation Extraction (Rosenman et al., 2020). In the more realistic end-to-end RE setting, we can expect yet another heuristic: the mere retention of training relation triples. In this paper we propose two experiments confirming that retention of known facts is a key factor of performance on standard benchmarks. Furthermore, one experiment suggests that a pipeline model able to use intermediate type representations is less prone to over-rely on retention.

Session 9D: Resources and Evaluation 4

Punta Cana 1,2 *Chair: Barbara Plank*

Automatic Text Evaluation through the Lens of Wasserstein Barycenters

Pierre Colombo et al. 14:30–14:45

A new metric `BaryScore` to evaluate text generation based on deep contextualized embeddings (*e.g.*, BERT, Roberta, ELMo) is introduced. This metric is motivated by a new framework relying on optimal transport tools, *i.e.*, Wasserstein distance and barycenter. By modelling the layer output of deep contextualized embeddings as a probability distribution rather than by a vector embedding; this framework provides a natural way to aggregate the different outputs through the Wasserstein space topology. In addition, it provides theoretical grounds to our metric and offers an alternative to available solutions (*e.g.*, MoverScore and BertScore). Numerical evaluation is performed on four different tasks: machine translation, summarization, data2text generation and image captioning. Our results show that `BaryScore` outperforms other BERT based metrics and exhibits more consistent behaviour in particular for text summarization.

Visually Grounded Reasoning across Languages and Cultures

Fangyu Liu et al. 14:45–15:00

The design of widespread vision-and-language datasets and pre-trained encoders directly adopts, or draws inspiration from, the concepts and images of ImageNet. While one can hardly overestimate how much this benchmark contributed to progress in computer vision, it is mostly derived from lexical databases and image queries in English, resulting in source material with a North American or Western European bias. Therefore, we devise a new protocol to construct an ImageNet-style hierarchy representative of more languages and cultures. In particular, we let the selection of both concepts and images be entirely driven by native speakers, rather than scraping them automatically. Specifically, we focus on a typologically diverse set of languages, namely, Indonesian, Mandarin Chinese, Swahili, Tamil, and Turkish. On top of the concepts and images obtained through this new protocol, we create a multilingual dataset for Multicultural Reasoning over Vision and Language (MaRVL) by eliciting statements from native speaker annotators about pairs of images. The task consists of discriminating whether each grounded statement is true or false. We establish a series of baselines using state-of-the-art models and find that their cross-lingual transfer performance lags dramatically behind supervised performance in English. These results invite us to reassess the robustness and accuracy of current state-of-the-art models beyond a narrow domain, but also open up new exciting challenges for the development of truly multilingual and multicultural systems.

Back to Square One: Artifact Detection, Training and Commonsense Disentanglement in the Winograd Schema

Yanai Elazar et al. 15:00–15:15

The Winograd Schema (WS) has been proposed as a test for measuring commonsense capabilities of models. Recently, pre-trained language model-based approaches have boosted performance on some WS benchmarks but the source of improvement is still not clear. This paper suggests that the apparent progress on WS may not necessarily reflect progress in commonsense reasoning. To support this claim, we first show that the current evaluation method of WS is sub-optimal and propose a modification that uses twin sentences for evaluation. We also propose two new baselines that indicate the existence of artifacts in WS benchmarks. We then develop a method for evaluating WS-like sentences in a zero-shot setting to account for the commonsense reasoning abilities acquired during the pretraining and observe that popular language models perform randomly in this setting when using our more strict evaluation. We conclude that the observed progress is mostly due to the use of supervision in training WS models, which is not likely to successfully support all the required commonsense reasoning skills and knowledge.

Robustness Evaluation of Entity Disambiguation Using Prior Probes: the Case of Entity Overshadowing

Vera Provatorova et al. 15:15–15:30

Entity disambiguation (ED) is the last step of entity linking (EL), when candidate entities are reranked according to the context they appear in. All datasets for training and evaluating models for EL consist of convenience samples, such as news articles and tweets, that propagate the prior probability bias of the entity distribution towards more frequently occurring entities. It was shown that the performance of the EL systems on such datasets is overestimated since it is possible to obtain higher accuracy scores by merely learning the prior. To provide a more adequate evaluation benchmark, we introduce the ShadowLink dataset, which includes 16K short text snippets annotated with entity mentions. We evaluate and report the performance of popular EL systems on the ShadowLink benchmark. The results show a considerable difference in accuracy between more and less common entities for all of the EL systems under evaluation, demonstrating the effect of prior probability bias and entity overshadowing.

IndoNLI: A Natural Language Inference Dataset for Indonesian

Rahmad Mahendra et al. 15:30–15:45

We present IndoNLI, the first human-elicited NLI dataset for Indonesian. We adapt the data collection protocol for MNLI and collect ~18K sentence pairs annotated by crowd workers and experts. The expert-annotated data is used exclusively as a test set. It is designed to provide a challenging test-bed for Indonesian NLI by explicitly incorporating various linguistic phenomena such as numerical reasoning, structural changes, idioms,

or temporal and spatial reasoning. Experiment results show that XLM-R outperforms other pre-trained models in our data. The best performance on the expert-annotated data is still far below human performance (13.4% accuracy gap), suggesting that this test set is especially challenging. Furthermore, our analysis shows that our expert-annotated data is more diverse and contains fewer annotation artifacts than the crowd-annotated data. We hope this dataset can help accelerate progress in Indonesian NLP research.

Agreeing to Disagree: Annotating Offensive Language Datasets with Annotators' Disagreement

Elisa Leonardelli et al. 15:45–16:00

Since state-of-the-art approaches to offensive language detection rely on supervised learning, it is crucial to quickly adapt them to the continuously evolving scenario of social media. While several approaches have been proposed to tackle the problem from an algorithmic perspective, so to reduce the need for annotated data, less attention has been paid to the quality of these data. Following a trend that has emerged recently, we focus on the level of agreement among annotators while selecting data to create offensive language datasets, a task involving a high level of subjectivity. Our study comprises the creation of three novel datasets of English tweets covering different topics and having five crowd-sourced judgments each. We also present an extensive set of experiments showing that selecting training and test data according to different levels of annotators' agreement has a strong effect on classifiers performance and robustness. Our findings are further validated in cross-domain experiments and studied using a popular benchmark dataset. We show that such hard cases, where low agreement is present, are not necessarily due to poor-quality annotation and we advocate for a higher presence of ambiguous cases in future datasets, in order to train more robust systems and better account for the different points of view expressed online.

Session 9E: Syntax

Punta Cana 3,4 *Chair: Natalie Schluter*

A Root of a Problem: Optimizing Single-Root Dependency Parsing

Miloš Stanojević and Shay B. Cohen 14:30–14:45

We describe two approaches to single-root dependency parsing that yield significant speed ups in such parsing. One approach has been previously used in dependency parsers in practice, but remains undocumented in the parsing literature, and is considered a heuristic. We show that this approach actually finds the optimal dependency tree. The second approach relies on simple reweighting of the inference graph being input to the dependency parser and has an optimal running time. Here, we again show that this approach is fully correct and identifies the highest-scoring parse tree. Our experiments demonstrate a manyfold speed up compared to a previous graph-based state-of-the-art parser without any loss in accuracy or optimality.

Efficient Sampling of Dependency Structure

Ran Zmigrod, Tim Vieira, and Ryan Cotterell 14:45–15:00

Probabilistic distributions over spanning trees in directed graphs are a fundamental model of dependency structure in natural language processing, syntactic dependency trees. In NLP, dependency trees often have an additional root constraint: only one edge may emanate from the root. However, no sampling algorithm has been presented in the literature to account for this additional constraint. In this paper, we adapt two spanning tree sampling algorithms to faithfully sample dependency trees from a graph subject to the root constraint. Wilson (1996('s sampling algorithm has a running time of O(H) where H is the mean hitting time of the graph. Colbourn (1996)'s sampling algorithm has a running time of $O(N^3)$, which is often greater than the mean hitting time of a directed graph. Additionally, we build upon Colbourn's algorithm and present a novel extension that can sample K trees without replacement in $O(K N^3 + K^2 N)$ time. To the best of our knowledge, no algorithm has been given for sampling spanning trees without replacement from a directed graph.

Efficient Computation of Expectations under Spanning Tree Distributions

Ran Zmigrod, Tim Vieira, and Ryan Cotterell 15:00–15:15

We give a general framework for inference in spanning tree models. We propose unified algorithms for the important cases of first-order expectations and second-order expectations in edge-factored, non-projective spanning-tree models. Our algorithms exploit a fundamental connection between gradients and expectations, which allows us to derive efficient algorithms. These algorithms are easy to implement with or without automatic differentiation software. We motivate the development of our framework with several cautionary tales of previous research, which has developed numerous inefficient algorithms for computing expectations and their gradients. We demonstrate how our framework efficiently computes several quantities with known algorithms, including the expected attachment score, entropy, and generalized expectation criteria. As a bonus, we give algorithms for quantities that are missing in the literature, including the KL divergence. In all cases, our approach matches the efficiency of existing algorithms and, in several cases, reduces the runtime complexity by a factor of the sentence length. We validate the implementation of our framework through runtime experiments. We find our algorithms are up to 15 and 9 times faster than previous algorithms for computing the Shannon entropy and the gradient of the generalized expectation objective, respectively.

Joint Universal Syntactic and Semantic Parsing

Elias Stengel-Eskin et al. 15:15–15:30

While numerous attempts have been made to jointly parse syntax and semantics, high performance in one domain typically comes at the price of performance in the other. This trade-off contradicts the large body of research focusing on the rich interactions at the syntax-semantics interface. We explore multiple model architectures which allow us to exploit the rich syntactic and semantic annotations contained in the Universal Decompositional Semantics (UDS) dataset, jointly parsing Universal Dependencies and UDS to obtain state-of-the-art results in both formalisms. We analyze the behaviour of a joint model of syntax and semantics, finding patterns supported by linguistic theory at the syntax-semantics interface. We then investigate to what degree joint modeling generalizes to a multilingual setting, where we find similar trends across 8 languages.

Reducing Discontinuous to Continuous Parsing with Pointer Network Reordering

Daniel Fernández-González and Carlos Gómez-Rodríguez 15:30–15:40

Discontinuous constituent parsers have always lagged behind continuous approaches in terms of accuracy and speed, as the presence of constituents with discontinuous yield introduces extra complexity to the task. However, a discontinuous tree can be converted into a continuous variant by reordering tokens. Based on that, we propose to reduce discontinuous parsing to a continuous problem, which can then be directly solved by any off-the-shelf continuous parser. To that end, we develop a Pointer Network capable of accurately generating the continuous token arrangement for a given input sentence and define a bijective function to recover the original order. Experiments on the main benchmarks with two continuous parsers prove that our approach is on par in accuracy with purely discontinuous state-of-the-art algorithms, but considerably faster.

A New Representation for Span-based CCG Parsing

Yoshihide Kato and Shigeki Matsubara 15:40–15:50

This paper proposes a new representation for CCG derivations. CCG derivations are represented as trees whose nodes are labeled with categories strictly restricted by CCG rule schemata. This characteristic is not suitable for span-based parsing models because they predict node labels independently. In other words, span-based models may generate invalid CCG derivations that violate the rule schemata. Our proposed representation decomposes CCG derivations into several independent pieces and prevents the span-based parsing models from violating the schemata. Our experimental result shows that an off-the-shelf span-based parser with our representation is comparable with previous CCG parsers.

Session 9F: Efficient Methods for NLP 3

Santo Domingo *Chair: Emma Strubell*

What to Pre-Train on? Efficient Intermediate Task Selection

Clifton Poth et al. 14:30–14:45

Intermediate task fine-tuning has been shown to culminate in large transfer gains across many NLP tasks. With an abundance of candidate datasets as well as pre-trained language models, it has become infeasible to experiment with all combinations to find the best transfer setting. In this work, we provide a comprehensive comparison of different methods for efficiently identifying beneficial tasks for intermediate transfer learning. We focus on parameter and computationally efficient adapter settings, highlight different data-availability scenarios, and provide expense estimates for each method. We experiment with a diverse set of 42 intermediate and 11 target English classification, multiple choice, question answering, and sequence tagging tasks. Our results demonstrate that efficient embedding based methods, which rely solely on the respective datasets, outperform computational expensive few-shot fine-tuning approaches. Our best methods achieve an average Regret3 of 1% across all target tasks, demonstrating that we are able to efficiently identify the best datasets for intermediate training.

PermuteFormer: Efficient Relative Position Encoding for Long Sequences

Peng Chen 14:45–15:00

A recent variation of Transformer, Performer, scales Transformer to longer sequences with a linear attention mechanism. However, it is not compatible with relative position encoding, which has advantages over absolute position encoding. In this paper, we discuss possible ways to add relative position encoding to Performer. Based on the analysis, we propose PermuteFormer, a Performer-based model with relative position encoding that scales linearly on long sequences. PermuteFormer applies position-dependent transformation on queries and keys to encode positional information into the attention module. This transformation is carefully crafted so that the final output of self-attention is not affected by absolute positions of tokens. PermuteFormer introduces negligible computational overhead by design that it runs as fast as Performer. We evaluate PermuteFormer on Long-Range Arena, a dataset for long sequences, as well as WikiText-103, a language modeling dataset. The experiments show that PermuteFormer uniformly improves the performance of Performer with almost no computational overhead and outperforms vanilla Transformer on most of the tasks.

Block Pruning For Faster Transformers

François Lagunas et al. 15:00–15:15

Pre-training has improved model accuracy for both classification and generation tasks at the cost of introducing much larger and slower models. Pruning methods have proven to be an effective way of reducing model size, whereas distillation methods are proven for speeding up inference. We introduce a block pruning approach targeting both small and fast models. Our approach extends structured methods by considering blocks of any size and integrates this structure into the movement pruning paradigm for fine-tuning. We find that this approach learns to prune out full components of the underlying model, such as attention heads. Experiments consider classification and generation tasks, yielding among other results a pruned model that is a 2.4x faster, 74% smaller BERT on SQuAD v1, with a 1% drop on F1, competitive both with distilled models in speed and pruned models in size.

Finetuning Pretrained Transformers into RNNs

Jungo Kasai et al. 15:15–15:30

Transformers have outperformed recurrent neural networks (RNNs) in natural language generation. But this comes with a signifi- cant computational cost, as the attention mechanism's complexity scales quadratically with sequence length. Efficient transformer variants have received increasing interest in recent works. Among them, a linear-complexity recurrent variant has proven well suited for autoregressive generation. It approximates the softmax attention with randomized or heuristic feature maps, but can be difficult to train and may yield suboptimal accuracy. This work aims to convert a pretrained transformer into its efficient recurrent counterpart, improving efficiency while maintaining accuracy. Specifically, we propose a swap-then-finetune procedure: in an off-the-shelf pretrained transformer, we replace the softmax attention with its linear-complexity recurrent alternative and then finetune. With a learned feature map, our approach provides an improved trade-off between efficiency and accuracy over the standard transformer and other recurrent variants. We also show that the finetuning process has lower training cost relative to training these recurrent variants from scratch. As many models for natural language tasks are increasingly dependent on large-scale pretrained transformers, this work presents a viable approach to improving inference efficiency without repeating the expensive pretraining process.

How to Train BERT with an Academic Budget

Peter Izsak, Moshe Berchansky, and Omer Levy 15:30–15:40

While large language models a la BERT are used ubiquitously in NLP, pretraining them is considered a luxury that only a few well-funded industry labs can afford. How can one train such models with a more modest budget? We present a recipe for pretraining a masked language model in 24 hours using a single low-end deep learning server. We demonstrate that through a combination of software optimizations, design choices, and hyperparameter tuning, it is possible to produce models that are competitive with BERT-base on GLUE tasks at a fraction of the original pretraining cost.

Beyond Preserved Accuracy: Evaluating Loyalty and Robustness of BERT Compression
Canwen Xu et al. 15:40–15:50

Recent studies on compression of pretrained language models (e.g., BERT) usually use preserved accuracy as the metric for evaluation. In this paper, we propose two new metrics, label loyalty and probability loyalty that measure how closely a compressed model (i.e., student) mimics the original model (i.e., teacher). We also explore the effect of compression with regard to robustness under adversarial attacks. We benchmark quantization, pruning, knowledge distillation and progressive module replacing with loyalty and robustness. By combining multiple compression techniques, we provide a practical strategy to achieve better accuracy, loyalty and robustness.

IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization
Fajri Koto, Jey Han Lau, and Timothy Baldwin 15:50–16:00

We present IndoBERTweet, the first large-scale pretrained model for Indonesian Twitter that is trained by extending a monolingually-trained Indonesian BERT model with additive domain-specific vocabulary. We focus in particular on efficient model adaptation under vocabulary mismatch, and benchmark different ways of initializing the BERT embedding layer for new word types. We find that initializing with the average BERT subword embedding makes pretraining five times faster, and is more effective than proposed methods for vocabulary adaptation in terms of extrinsic evaluation over seven Twitter-based datasets.

Poster and Demo session 6

In-person Poster Session 6

Bavaro 4 *14:30–16:00*

Types of Out-of-Distribution Texts and How to Detect Them

Udit Arora, William Huang, and He He

Despite agreement on the importance of detecting out-of-distribution (OOD) examples, there is little consensus on the formal definition of the distribution shifts of OOD examples and how to best detect them. We categorize these examples as exhibiting a background shift or semantic shift, and find that the two major approaches to OOD detection, calibration and density estimation (language modeling for text), have distinct behavior on these types of OOD data. Across 14 pairs of in-distribution and OOD English natural language understanding datasets, we find that density estimation methods consistently beat calibration methods in background shift settings and perform worse in semantic shift settings. In addition, we find that both methods generally fail to detect examples from challenge data, indicating that these examples constitute a different type of OOD data. Overall, while the categorization we apply explains many of the differences between the two methods, our results call for a more explicit definition of OOD to create better benchmarks and build detectors that can target the type of OOD data expected at test time.

Self-training with Few-shot Rationalization

Meghana Moorthy Bhat, Alessandro Sordoni, and Subhabrata Mukherjee

While pre-trained language models have obtained state-of-the-art performance for several natural language understanding tasks, they are quite opaque in terms of their decision-making process. While some recent works focus on rationalizing neural predictions by highlighting salient concepts in the text as justifications or rationales, they rely on thousands of labeled training examples for both task labels as well as annotated rationales for every instance. Such extensive large-scale annotations are infeasible to obtain for many tasks. To this end, we develop a multi-task teacher-student framework based on self-training pre-trained language models with limited task-specific labels and rationales and judicious sample selection to learn from informative pseudo-labeled examples. We study several characteristics of what constitutes a good rationale and demonstrate that the neural model performance can be significantly improved by making it aware of its rationalized predictions, particularly in low-resource settings. Extensive experiments in several benchmark datasets demonstrate the effectiveness of our approach.

MTAdam: Automatic Balancing of Multiple Training Loss Terms

Itzik Malkiel and Lior Wolf

When training neural models, it is common to combine multiple loss terms. The balancing of these terms requires considerable human effort and is computationally demanding. Moreover, the optimal trade-off between the loss terms can change as training progresses, e.g., for adversarial terms. In this work, we generalize the Adam optimization algorithm to handle multiple loss terms. The guiding principle is that for every layer, the gradient magnitude of the terms should be balanced. To this end, the Multi-Term Adam (MTAdam) computes the derivative of each loss term separately, infers the first and second moments per parameter and loss term, and calculates a first moment for the magnitude per layer of the gradients arising from each loss. This magnitude is used to continuously balance the gradients across all layers, in a manner that both varies from one layer to the next and dynamically changes over time. Our results show that training with the new method leads to fast recovery from suboptimal initial loss weighting and to training outcomes that match or improve conventional training with the prescribed hyperparameters of each method.

Softmax Tree: An Accurate, Fast Classifier When the Number of Classes Is Large

Arman Zharmagambetov, Magzhan Gabidolla, and Miguel A. Carreira-Perpinan

Classification problems having thousands or more classes naturally occur in NLP, for example language models or document classification. A softmax or one-vs-all classifier naturally handles many classes, but it is very slow at inference time, because every class score must be calculated to find the top class. We propose the "softmax tree", consisting of a binary tree having sparse hyperplanes at the decision nodes (which make hard, not soft, decisions) and small softmax classifiers at the leaves. This is much faster at inference because the input instance follows a single path to a leaf (whose length is logarithmic on the number of leaves) and the softmax classifier at each leaf operates on a small subset of the classes. Although learning accurate tree-based models has proven difficult in the past, we are able to overcome this by using a variation of a recent algorithm, tree alternating optimization (TAO). Compared to a softmax and other classifiers, the resulting softmax trees are both more accurate in prediction and faster in inference, as shown in NLP problems having from one thousand to one hundred thousand classes.

Multivalent Entailment Graphs for Question Answering

Nick McKenna et al.

Drawing inferences between open-domain natural language predicates is a necessity for true language understanding. There has been much progress in unsupervised learning of entailment graphs for this purpose. We make three contributions: (1) we reinterpret the Distributional Inclusion Hypothesis to model entailment between predicates of different valencies, like DEFEAT(Biden, Trump) entails WIN(Biden); (2) we actualize

this theory by learning unsupervised Multivalent Entailment Graphs of open-domain predicates; and (3) we demonstrate the capabilities of these graphs on a novel question answering task. We show that directional entailment is more helpful for inference than non-directional similarity on questions of fine-grained semantics. We also show that drawing on evidence across valencies answers more questions than by using only the same valency evidence.

Is Everything in Order? A Simple Way to Order Sentences

Somnath Basu Roy Chowdhury, Faeze Brahman, and Snigdha Chaturvedi

The task of organizing a shuffled set of sentences into a coherent text has been used to evaluate a machine's understanding of causal and temporal relations. We formulate the sentence ordering task as a conditional text-to-marker generation problem. We present Reorder-BART (Re-BART) that leverages a pre-trained Transformer-based model to identify a coherent order for a given set of shuffled sentences. The model takes a set of shuffled sentences with sentence-specific markers as input and generates a sequence of position markers of the sentences in the ordered text. Re-BART achieves the state-of-the-art performance across 7 datasets in Perfect Match Ratio (PMR) and Kendall's tau. We perform evaluations in a zero-shot setting, showcasing that our model is able to generalize well across other datasets. We additionally perform several experiments to understand the functioning and limitations of our framework.

VeeAlign: Multifaceted Context Representation Using Dual Attention for Ontology Alignment

Vivek Iyer, Arvind Agarwal, and Harshit Kumar

Ontology Alignment is an important research problem applied to various fields such as data integration, data transfer, data preparation, etc. State-of-the-art (SOTA) Ontology Alignment systems typically use naive domain-dependent approaches with handcrafted rules or domain-specific architectures, making them unscalable and inefficient. In this work, we propose VeeAlign, a Deep Learning based model that uses a novel dual-attention mechanism to compute the contextualized representation of a concept which, in turn, is used to discover alignments. By doing this, not only is our approach able to exploit both syntactic and semantic information encoded in ontologies, it is also, by design, flexible and scalable to different domains with minimal effort. We evaluate our model on four different datasets from different domains and languages, and establish its superiority through these results as well as detailed ablation studies. The code and datasets used are available at https://github.com/Remorax/VeeAlign.

Finding needles in a haystack: Sampling Structurally-diverse Training Sets from Synthetic Data for Compositional Generalization

Inbar Oren, Jonathan Herzig, and Jonathan Berant

Modern semantic parsers suffer from two principal limitations. First, training requires expensive collection of utterance-program pairs. Second, semantic parsers fail to generalize at test time to new compositions/structures that have not been observed during training. Recent research has shown that automatic generation of synthetic utterance-program pairs can alleviate the first problem, but its potential for the second has thus far been underexplored. In this work, we investigate automatic generation of synthetic utterance-program pairs for improving compositional generalization in semantic parsing. Given a small training set of annotated examples and an "infinite" pool of synthetic examples, we select a subset of synthetic examples that are structurally-diverse and use them to improve compositional generalization. We evaluate our approach on a new split of the schema2QA dataset, and show that it leads to dramatic improvements in compositional generalization as well as moderate improvements in the traditional i.i.d setup. Moreover, structurally-diverse sampling achieves these improvements with as few as 5K examples, compared to 1M examples when sampling uniformly at random – a 200x improvement in data efficiency.

GeneSis: A Generative Approach to Substitutes in Context

Caterina Lacerra, Rocco Tripodi, and Roberto Navigli

The lexical substitution task aims at generating a list of suitable replacements for a target word in context, ideally keeping the meaning of the modified text unchanged. While its usage has increased in recent years, the paucity of annotated data prevents the finetuning of neural models on the task, hindering the full fruition of recently introduced powerful architectures such as language models. Furthermore, lexical substitution is usually evaluated in a framework that is strictly bound to a limited vocabulary, making it impossible to credit appropriate, but out-of-vocabulary, substitutes. To assess these issues, we proposed GeneSis (Generating Substitutes in contexts), the first generative approach to lexical substitution. Thanks to a seq2seq model, we generate substitutes for a word according to the context it appears in, attaining state-of-the-art results on different benchmarks. Moreover, our approach allows silver data to be produced for further improving the performances of lexical substitution systems. Along with an extensive analysis of GeneSis results, we also present a human evaluation of the generated substitutes in order to assess their quality. We release the fine-tuned models, the generated datasets, and the code to reproduce the experiments at https://github.com/SapienzaNLP/genesis.

Semi-Supervised Exaggeration Detection of Health Science Press Releases

Dustin Wright and Isabelle Augenstein

Public trust in science depends on honest and factual communication of scientific papers. However, recent studies have demonstrated a tendency of news media to misrepresent scientific papers by exaggerating their findings. Given this, we present a formalization of and study into the problem of exaggeration detection in

science communication. While there are an abundance of scientific papers and popular media articles written about them, very rarely do the articles include a direct link to the original paper, making data collection challenging, and necessitating the need for few-shot learning. We address this by curating a set of labeled press release/abstract pairs from existing expert annotated studies on exaggeration in press releases of scientific papers suitable for benchmarking the performance of machine learning models on the task. Using limited data from this and previous studies on exaggeration detection in science, we introduce MT-PET, a multi-task version of Pattern Exploiting Training (PET), which leverages knowledge from complementary cloze-style QA tasks to improve few-shot learning. We demonstrate that MT-PET outperforms PET and supervised learning both when data is limited, as well as when there is an abundance of data for the main task.

Phrase-BERT: Improved Phrase Embeddings from BERT with an Application to Corpus Exploration

Shufan Wang, Laure Thompson, and Mohit Iyyer

Phrase representations derived from BERT often do not exhibit complex phrasal compositionality, as the model relies instead on lexical similarity to determine semantic relatedness. In this paper, we propose a contrastive fine-tuning objective that enables BERT to produce more powerful phrase embeddings. Our approach (Phrase-BERT) relies on a dataset of diverse phrasal paraphrases, which is automatically generated using a paraphrase generation model, as well as a large-scale dataset of phrases in context mined from the Books3 corpus. Phrase-BERT outperforms baselines across a variety of phrase-level similarity tasks, while also demonstrating increased lexical diversity between nearest neighbors in the vector space. Finally, as a case study, we show that Phrase-BERT embeddings can be easily integrated with a simple autoencoder to build a phrase-based neural topic model that interprets topics as mixtures of words and phrases by performing a nearest neighbor search in the embedding space. Crowdsourced evaluations demonstrate that this phrase-based topic model produces more coherent and meaningful topics than baseline word and phrase-level topic models, further validating the utility of Phrase-BERT.

Detecting Contact-Induced Semantic Shifts: What Can Embedding-Based Methods Do in Practice?

Filip Miletic, Anne Przewozny-Desriaux, and Ludovic Tanguy

This study investigates the applicability of semantic change detection methods in descriptively oriented linguistic research. It specifically focuses on contact-induced semantic shifts in Quebec English. We contrast synchronic data from different regions in order to identify the meanings that are specific to Quebec and potentially related to language contact. Type-level embeddings are used to detect new semantic shifts, and token-level embeddings to isolate regionally specific occurrences. We introduce a new 80-item test set and conduct both quantitative and qualitative evaluations. We demonstrate that diachronic word embedding methods can be applied to contact-induced semantic shifts observed in synchrony, obtaining results comparable to the state of the art on similar tasks in diachrony. However, we show that encouraging evaluation results do not translate to practical value in detecting new semantic shifts. Finally, our application of token-level embeddings accelerates manual data exploration and provides an efficient way of scaling up sociolinguistic analyses.

Causal Direction of Data Collection Matters: Implications of Causal and Anticausal Learning for NLP

Zhijing Jin et al.

The principle of independent causal mechanisms (ICM) states that generative processes of real world data consist of independent modules which do not influence or inform each other. While this idea has led to fruitful developments in the field of causal inference, it is not widely-known in the NLP community. In this work, we argue that the causal direction of the data collection process bears nontrivial implications that can explain a number of published NLP findings, such as differences in semi-supervised learning (SSL) and domain adaptation (DA) performance across different settings. We categorize common NLP tasks according to their causal direction and empirically assay the validity of the ICM principle for text data using minimum description length. We conduct an extensive meta-analysis of over 100 published SSL and 30 DA studies, and find that the results are consistent with our expectations based on causal insights. This work presents the first attempt to analyze the ICM principle in NLP, and provides constructive suggestions for future modeling choices.

Effects of Parameter Norm Growth During Transformer Training: Inductive Bias from Gradient Descent

William Merrill et al.

The capacity of neural networks like the widely adopted transformer is known to be very high. Evidence is emerging that they learn successfully due to inductive bias in the training routine, typically a variant of gradient descent (GD). To better understand this bias, we study the tendency for transformer parameters to grow in magnitude (ℓ_2 norm) during training, and its implications for the emergent representations within self attention layers. Empirically, we document norm growth in the training of transformer language models, including T5 during its pretraining. As the parameters grow in magnitude, we prove that the network approximates a discretized network with saturated activation functions. Such "saturated" networks are known to have a reduced capacity compared to the full network family that can be described in terms of formal languages and automata. Our results suggest saturation is a new characterization of an inductive bias implicit in GD of particular interest for NLP. We leverage the emergent discrete structure in a saturated transformer to analyze the role of different attention heads, finding that some focus locally on a small number of positions, while other

heads compute global averages, allowing counting. We believe understanding the interplay between these two capabilities may shed further light on the structure of computation within large transformers.

The Devil is in the Detail: Simple Tricks Improve Systematic Generalization of Transformers

Róbert Csordás, Kazuki Irie, and Juergen Schmidhuber

Recently, many datasets have been proposed to test the systematic generalization ability of neural networks. The companion baseline Transformers, typically trained with default hyper-parameters from standard tasks, are shown to fail dramatically. Here we demonstrate that by revisiting model configurations as basic as scaling of embeddings, early stopping, relative positional embedding, and Universal Transformer variants, we can drastically improve the performance of Transformers on systematic generalization. We report improvements on five popular datasets: SCAN, CFQ, PCFG, COGS, and Mathematics dataset. Our models improve accuracy from 50% to 85% on the PCFG productivity split, and from 35% to 81% on COGS. On SCAN, relative positional embedding largely mitigates the EOS decision problem (Newman et al., 2020), yielding 100% accuracy on the length split with a cutoff at 26. Importantly, performance differences between these models are typically invisible on the IID data split. This calls for proper generalization validation sets for developing neural networks that generalize systematically. We publicly release the code to reproduce our results.

Contrastive Out-of-Distribution Detection for Pretrained Transformers

Wenxuan Zhou, Fangyu Liu, and Muhao Chen

Pretrained Transformers achieve remarkable performance when training and test data are from the same distribution. However, in real-world scenarios, the model often faces out-of-distribution (OOD) instances that can cause severe semantic shift problems at inference time. Therefore, in practice, a reliable model should identify such instances, and then either reject them during inference or pass them over to models that handle another distribution. In this paper, we develop an unsupervised OOD detection method, in which only the in-distribution (ID) data are used in training. We propose to fine-tune the Transformers with a contrastive loss, which improves the compactness of representations, such that OOD instances can be better differentiated from ID ones. These OOD instances can then be accurately detected using the Mahalanobis distance in the model's penultimate layer. We experiment with comprehensive settings and achieve near-perfect OOD detection performance, outperforming baselines drastically. We further investigate the rationales behind the improvement, finding that more compact representations through margin-based contrastive learning bring the improvement. We release our code to the community for future research.

Competency Problems: On Finding and Removing Artifacts in Language Data

Matt Gardner et al.

Much recent work in NLP has documented dataset artifacts, bias, and spurious correlations between input features and output labels. However, how to tell which features have "spurious" instead of legitimate correlations is typically left unspecified. In this work we argue that for complex language understanding tasks, all simple feature correlations are spurious, and we formalize this notion into a class of problems which we call competency problems. For example, the word "amazing" on its own should not give information about a sentiment label independent of the context in which it appears, which could include negation, metaphor, sarcasm, etc. We theoretically analyze the difficulty of creating data for competency problems when human bias is taken into account, showing that realistic datasets will increasingly deviate from competency problems as dataset size increases. This analysis gives us a simple statistical test for dataset artifacts, which we use to show more subtle biases than were described in prior work, including demonstrating that models are inappropriately affected by these less extreme biases. Our theoretical treatment of this problem also allows us to analyze proposed solutions, such as making local edits to dataset instances, and to give recommendations for future data collection and model design efforts that target competency problems.

Minimal Supervision for Morphological Inflection

Omer Goldman and Reut Tsarfaty

Neural models for the various flavours of morphological reinflection tasks have proven to be extremely accurate given ample labeled data, yet labeled data may be slow and costly to obtain. In this work we aim to overcome this annotation bottleneck by bootstrapping labeled data from a seed as small as *five* labeled inflection tables, accompanied by a large bulk of unlabeled text. Our bootstrapping method exploits the orthographic and semantic regularities in morphological systems in a two-phased setup, where word tagging based on *analogies* is followed by word pairing based on *distances*. Our experiments with the Paradigm Cell Filling Problem over eight typologically different languages show that in languages with relatively simple morphology, orthographic regularities on their own allow inflection models to achieve respectable accuracy. Combined orthographic and semantic regularities alleviate difficulties with particularly complex morpho-phonological systems. We further show that our bootstrapping methods substantially outperform hallucination-based methods commonly used for overcoming the annotation bottleneck in morphological reinflection tasks.

Cross-Domain Label-Adaptive Stance Detection

Momchil Hardalov et al.

Stance detection concerns the classification of a writer's viewpoint towards a target. There are different task variants, e.g., stance of a tweet vs. a full article, or stance with respect to a claim vs. an (implicit) topic. Moreover, task definitions vary, which includes the label inventory, the data collection, and the annotation

protocol. All these aspects hinder cross-domain studies, as they require changes to standard domain adaptation approaches. In this paper, we perform an in-depth analysis of 16 stance detection datasets, and we explore the possibility for cross-domain learning from them. Moreover, we propose an end-to-end unsupervised framework for out-of-domain prediction of unseen, user-defined labels. In particular, we combine domain adaptation techniques such as mixture of experts and domain-adversarial training with label embeddings, and we demonstrate sizable performance gains over strong baselines, both (i) in-domain, i.e., for seen targets, and (ii) out-of-domain, i.e., for unseen targets. Finally, we perform an exhaustive analysis of the cross-domain results, and we highlight the important factors influencing the model performance.

Fast, Effective, and Self-Supervised: Transforming Masked Language Models into Universal Lexical and Sentence Encoders
Fangyu Liu et al.

Previous work has indicated that pretrained Masked Language Models (MLMs) are not effective as universal lexical and sentence encoders off-the-shelf, i.e., without further task-specific fine-tuning on NLI, sentence similarity, or paraphrasing tasks using annotated task data. In this work, we demonstrate that it is possible to turn MLMs into effective lexical and sentence encoders even without any additional data, relying simply on self-supervision. We propose an extremely simple, fast, and effective contrastive learning technique, termed Mirror-BERT, which converts MLMs (e.g., BERT and RoBERTa) into such encoders in 20-30 seconds with no access to additional external knowledge. Mirror-BERT relies on identical and slightly modified string pairs as positive (i.e., synonymous) fine-tuning examples, and aims to maximise their similarity during "identity fine-tuning". We report huge gains over off-the-shelf MLMs with Mirror-BERT both in lexical-level and in sentence-level tasks, across different domains and different languages. Notably, in sentence similarity (STS) and question-answer entailment (QNLI) tasks, our self-supervised Mirror-BERT model even matches the performance of the Sentence-BERT models from prior work which rely on annotated task data. Finally, we delve deeper into the inner workings of MLMs, and suggest some evidence on why this simple Mirror-BERT fine-tuning approach can yield effective universal lexical and sentence encoders.

Implicit Premise Generation with Discourse-aware Commonsense Knowledge Models
Tuhin Chakrabarty, Aadit Trivedi, and Smaranda Muresan

Enthymemes are defined as arguments where a premise or conclusion is left implicit. We tackle the task of generating the *implicit premise in an enthymeme*, which requires not only an understanding of the stated conclusion and premise but also additional inferences that could depend on commonsense knowledge. The largest available dataset for enthymemes (Habernal et al., 2018) consists of 1.7k samples, which is not large enough to train a neural text generation model. To address this issue, we take advantage of a similar task and dataset: Abductive reasoning in narrative text (Bhagavatula et al., 2020). However, we show that simply using a state-of-the-art seq2seq model fine-tuned on this data might not generate meaningful implicit premises associated with the given enthymemes. We demonstrate that encoding discourse-aware commonsense during fine-tuning improves the quality of the generated implicit premises and outperforms all other baselines both in automatic and human evaluations on three different datasets.

RuleBERT: Teaching Soft Rules to Pre-Trained Language Models
Mohammed Saeed et al.

While pre-trained language models (PLMs) are the go-to solution to tackle many natural language processing problems, they are still very limited in their ability to capture and to use common-sense knowledge. In fact, even if information is available in the form of approximate (soft) logical rules, it is not clear how to transfer it to a PLM in order to improve its performance for deductive reasoning tasks. Here, we aim to bridge this gap by teaching PLMs how to reason with soft Horn rules. We introduce a classification task where, given facts and soft rules, the PLM should return a prediction with a probability for a given hypothesis. We release the first dataset for this task, and we propose a revised loss function that enables the PLM to learn how to predict precise probabilities for the task. Our evaluation results show that the resulting fine-tuned models achieve very high performance, even on logical rules that were unseen at training. Moreover, we demonstrate that logical notions expressed by the rules are transferred to the fine-tuned model, yielding state-of-the-art results on external datasets.

Asking It All: Generating Contextualized Questions for any Semantic Role
Valentina Pyatkin et al.

Asking questions about a situation is an inherent step towards understanding it. To this end, we introduce the task of role question generation, which, given a predicate mention and a passage, requires producing a set of questions asking about all possible semantic roles of the predicate. We develop a two-stage model for this task, which first produces a context-independent question prototype for each role and then revises it to be contextually appropriate for the passage. Unlike most existing approaches to question generation, our approach does not require conditioning on existing answers in the text. Instead, we condition on the type of information to inquire about, regardless of whether the answer appears explicitly in the text, could be inferred from it, or should be sought elsewhere. Our evaluation demonstrates that we generate diverse and well-formed questions for a large, broad-coverage ontology of predicates and roles.

ConSeC: Word Sense Disambiguation as Continuous Sense Comprehension
Edoardo Barba, Luigi Procopio, and Roberto Navigli

Supervised systems have nowadays become the standard recipe for Word Sense Disambiguation (WSD), with Transformer-based language models as their primary ingredient. However, while these systems have certainly attained unprecedented performances, virtually all of them operate under the constraining assumption that, given a context, each word can be disambiguated individually with no account of the other sense choices. To address this limitation and drop this assumption, we propose CONtinuous SEnse Comprehension (ConSeC), a novel approach to WSD: leveraging a recent re-framing of this task as a text extraction problem, we adapt it to our formulation and introduce a feedback loop strategy that allows the disambiguation of a target word to be conditioned not only on its context but also on the explicit senses assigned to nearby words. We evaluate ConSeC and examine how its components lead it to surpass all its competitors and set a new state of the art on English WSD. We also explore how ConSeC fares in the cross-lingual setting, focusing on 8 languages with various degrees of resource availability, and report significant improvements over prior systems. We release our code at https://github.com/SapienzaNLP/consec.

Numeracy enhances the Literacy of Language Models
Avijit Thawani, Jay Pujara, and Filip Ilievski

Specialized number representations in NLP have shown improvements on numerical reasoning tasks like arithmetic word problems and masked number prediction. But humans also use numeracy to make better sense of world concepts, e.g., you can seat 5 people in your 'room' but not 500. Does a better grasp of numbers improve a model's understanding of other concepts and words? This paper studies the effect of using six different number encoders on the task of masked word prediction (MWP), as a proxy for evaluating literacy. To support this investigation, we develop Wiki-Convert, a 900,000 sentence dataset annotated with numbers and units, to avoid conflating nominal and ordinal number occurrences. We find a significant improvement in MWP for sentences containing numbers, that exponent embeddings are the best number encoders, yielding over 2 points jump in prediction accuracy over a BERT baseline, and that these enhanced literacy skills also generalize to contexts without annotated numbers. We release all code at https://git.io/JuZXn.

SPARQLing Database Queries from Intermediate Question Decompositions
Irina Saparina and Anton Osokin

To translate natural language questions into executable database queries, most approaches rely on a fully annotated training set. Annotating a large dataset with queries is difficult as it requires query-language expertise. We reduce this burden using grounded in databases intermediate question representations. These representations are simpler to collect and were originally crowdsourced within the Break dataset (Wolfson et al., 2020). Our pipeline consists of two parts: a neural semantic parser that converts natural language questions into the intermediate representations and a non-trainable transpiler to the SPARQL query language (a standard language for accessing knowledge graphs and semantic web). We chose SPARQL because its queries are structurally closer to our intermediate representations (compared to SQL). We observe that the execution accuracy of queries constructed by our model on the challenging Spider dataset is comparable with the state-of-the-art text-to-SQL methods trained with annotated SQL queries. Our code and data are publicly available (https://github.com/yandex-research/sparqling-queries).

In-person Demo Session

Bavaro 4 *14:30–16:00*

AMuSE-WSD: An All-in-one Multilingual System for Easy Word Sense Disambiguation
Riccardo Orlando et al.

Over the past few years, Word Sense Disambiguation (WSD) has received renewed interest: recently proposed systems have shown the remarkable effectiveness of deep learning techniques in this task, especially when aided by modern pretrained language models. Unfortunately, such systems are still not available as ready-to-use end-to-end packages, making it difficult for researchers to take advantage of their performance. The only alternative for a user interested in applying WSD to downstream tasks is to rely on currently available end-to-end WSD systems, which, however, still rely on graph-based heuristics or non-neural machine learning algorithms. In this paper, we fill this gap and propose AMuSE-WSD, the first end-to-end system to offer high-quality sense information in 40 languages through a state-of-the-art neural model for WSD. We hope that AMuSE-WSD will provide a stepping stone for the integration of meaning into real-world applications and encourage further studies in lexical semantics. AMuSE-WSD is available online at http://nlp.uniroma1.it/amuse-wsd.

iFacetSum: Coreference-based Interactive Faceted Summarization for Multi-Document Exploration
Eran Hirsch et al.

We introduce iFacetSum, a web application for exploring topical document collections. iFacetSum integrates interactive summarization together with faceted search, by providing a novel faceted navigation scheme that yields abstractive summaries for the user selections. This approach offers both a comprehensive overview as well as particular details regarding subtopics of choice. The facets are automatically produced based on cross-document coreference pipelines, rendering generic concepts, entities and statements surfacing in the source texts. We analyze the effectiveness of our application through small-scale user studies that suggest the usefulness of our tool.

Virtual Poster and Demo Session 3

Time: *8:30–10:30*

Classifying Dyads for Militarized Conflict Analysis
Niklas Stoehr et al.

Understanding the origins of militarized conflict is a complex, yet important undertaking. Existing research seeks to build this understanding by considering bi-lateral relationships between entity pairs (dyadic causes) and multi-lateral relationships among multiple entities (systemic causes). The aim of this work is to compare these two causes in terms of how they correlate with conflict between two entities. We do this by devising a set of textual and graph-based features which represent each of the causes. The features are extracted from Wikipedia and modeled as a large graph. Nodes in this graph represent entities connected by labeled edges representing ally or enemy-relationships. This allows casting the problem as an edge classification task, which we term dyad classification. We propose and evaluate classifiers to determine if a particular pair of entities are allies or enemies. Our results suggest that our systemic features might be slightly better correlates of conflict. Further, we find that Wikipedia articles of allies are semantically more similar than enemies.

Point-of-Interest Type Prediction using Text and Images
Danae Sánchez Villegas and Nikolaos Aletras

Point-of-interest (POI) type prediction is the task of inferring the type of a place from where a social media post was shared. Inferring a POI's type is useful for studies in computational social science including sociolinguistics, geosemiotics, and cultural geography, and has applications in geosocial networking technologies such as recommendation and visualization systems. Prior efforts in POI type prediction focus solely on text, without taking visual information into account. However in reality, the variety of modalities, as well as their semiotic relationships with one another, shape communication and interactions in social media. This paper presents a study on POI type prediction using multimodal information from text and images available at posting time. For that purpose, we enrich a currently available data set for POI type prediction with the images that accompany the text messages. Our proposed method extracts relevant information from each modality to effectively capture interactions between text and image achieving a macro F1 of 47.21 across 8 categories significantly outperforming the state-of-the-art method for POI type prediction based on text-only methods. Finally, we provide a detailed analysis to shed light on cross-modal interactions and the limitations of our best performing model.

Come hither or go away? Recognising pre-electoral coalition signals in the news
Ines Rehbein et al.

In this paper, we introduce the task of political coalition signal prediction from text, that is, the task of recognizing from the news coverage leading up to an election the (un)willingness of political parties to form a government coalition. We decompose our problem into two related, but distinct tasks: (i) predicting whether a reported statement from a politician or a journalist refers to a potential coalition and (ii) predicting the polarity of the signal – namely, whether the speaker is in favour of or against the coalition. For this, we explore the benefits of multi-task learning and investigate which setup and task formulation is best suited for each sub-task. We evaluate our approach, based on hand-coded newspaper articles, covering elections in three countries (Ireland, Germany, Austria) and two languages (English, German). Our results show that the multi-task learning approach can further improve results over a strong monolingual transfer learning baseline.

#HowYouTagTweets: Learning User Hashtagging Preferences via Personalized Topic Attention
Yuji Zhang et al.

Millions of hashtags are created on social media every day to cross-refer messages concerning similar topics. To help people find the topics they want to discuss, this paper characterizes a user's hashtagging preferences via predicting how likely they will post with a hashtag. It is hypothesized that one's interests in a hashtag are related with what they said before (user history) and the existing posts present the hashtag (hashtag contexts). These factors are married in the deep semantic space built with a pre-trained BERT and a neural topic model via multitask learning. In this way, user interests learned from the past can be customized to match future hashtags, which is beyond the capability of existing methods assuming unchanged hashtag semantics. Furthermore, we propose a novel personalized topic attention to capture salient contents to personalize hashtag contexts. Experiments on a large-scale Twitter dataset show that our model significantly outperforms the state-of-the-art recommendation approach without exploiting latent topics.

Learning Neural Templates for Recommender Dialogue System
Zujie Liang et al.

The task of Conversational Recommendation System (CRS), i.e., recommender dialog system, aims to recommend precise items to users through natural language interactions. Though recent end-to-end neural models have shown promising progress on this task, two key challenges still remain. First, the recommended items cannot be always incorporated into the generated response precisely and appropriately. Second, only the items mentioned in the training corpus have a chance to be recommended in the conversation. To tackle these challenges, we introduce a novel framework called NTRD for recommender dialogue system that can

decouple the dialogue generation from the item recommendation. NTRD has two key components, i.e., response template generator and item selector. The former adopts an encoder-decoder model to generate a response template with slot locations tied to target items, while the latter fills in slot locations with the proper items using a sufficient attention mechanism. Our approach combines the strengths of both classical slot filling approaches (that are generally controllable) and modern neural NLG approaches (that are generally more natural and accurate). Extensive experiments on the benchmark ReDial show our approach significantly outperforms the previous state-of-the-art methods. Besides, our approach has the unique advantage to produce novel items that do not appear in the training set of dialogue corpus. The code is available at `https://github.com/jokieleung/NTRD`.

Proxy Indicators for the Quality of Open-domain Dialogues
Rostislav Nedelchev, Jens Lehmann, and Ricardo Usbeck

The automatic evaluation of open-domain dialogues remains a largely unsolved challenge. Despite the abundance of work done in the field, human judges have to evaluate dialogues' quality. As a consequence, performing such evaluations at scale is usually expensive. This work investigates using a deep-learning model trained on the General Language Understanding Evaluation (GLUE) benchmark to serve as a quality indication of open-domain dialogues. The aim is to use the various GLUE tasks as different perspectives on judging the quality of conversation, thus reducing the need for additional training data or responses that serve as quality references. Due to this nature, the method can infer various quality metrics and can derive a component-based overall score. We achieve statistically significant correlation coefficients of up to 0.7.

Q2: Evaluating Factual Consistency in Knowledge-Grounded Dialogues via Question Generation and Question Answering
Or Honovich et al.

Neural knowledge-grounded generative models for dialogue often produce content that is factually inconsistent with the knowledge they rely on, making them unreliable and limiting their applicability. Inspired by recent work on evaluating factual consistency in abstractive summarization, we propose an automatic evaluation metric for factual consistency in knowledge-grounded dialogue using automatic question generation and question answering. Our metric, denoted Q^2, compares answer spans using natural language inference (NLI), instead of token-based matching as done in previous work. To foster proper evaluation, we curate a novel dataset of dialogue system outputs for the Wizard-of-Wikipedia dataset, manually annotated for factual consistency. We perform a thorough meta-evaluation of Q^2 against other metrics using this dataset and two others, where it consistently shows higher correlation with human judgements.

Knowledge-Aware Graph-Enhanced GPT-2 for Dialogue State Tracking
Weizhe Lin, Bo-Hsiang Tseng, and Bill Byrne

Dialogue State Tracking is central to multi-domain task-oriented dialogue systems, responsible for extracting information from user utterances. We present a novel hybrid architecture that augments GPT-2 with representations derived from Graph Attention Networks in such a way to allow causal, sequential prediction of slot values. The model architecture captures inter-slot relationships and dependencies across domains that otherwise can be lost in sequential prediction. We report improvements in state tracking performance in MultiWOZ 2.0 against a strong GPT-2 baseline and investigate a simplified sparse training scenario in which DST models are trained only on session-level annotations but evaluated at the turn level. We further report detailed analyses to demonstrate the effectiveness of graph models in DST by showing that the proposed graph modules capture inter-slot dependencies and improve the predictions of values that are common to multiple domains.

A Collaborative Multi-agent Reinforcement Learning Framework for Dialog Action Decomposition
Huimin Wang and Kam-Fai Wong

Most reinforcement learning methods for dialog policy learning train a centralized agent that selects a predefined joint action concatenating domain name, intent type, and slot name. The centralized dialog agent suffers from a great many user-agent interaction requirements due to the large action space. Besides, designing the concatenated actions is laborious to engineers and maybe struggled with edge cases. To solve these problems, we model the dialog policy learning problem with a novel multi-agent framework, in which each part of the action is led by a different agent. The framework reduces labor costs for action templates and decreases the size of the action space for each agent. Furthermore, we relieve the non-stationary problem caused by the changing dynamics of the environment as evolving of agents' policies by introducing a joint optimization process that makes agents can exchange their policy information. Concurrently, an independent experience replay buffer mechanism is integrated to reduce the dependence between gradients of samples to improve training efficiency. The effectiveness of the proposed framework is demonstrated in a multi-domain environment with both user simulator evaluation and human evaluation.

Zero-Shot Dialogue State Tracking via Cross-Task Transfer
Zhaojiang Lin et al.

Zero-shot transfer learning for dialogue state tracking (DST) enables us to handle a variety of task-oriented dialogue domains without the expense of collecting in-domain data. In this work, we propose to transfer the cross-task knowledge from general question answering (QA) corpora for the zero-shot DST task. Specifically, we propose TransferQA, a transferable generative QA model that seamlessly combines extractive QA

and multi-choice QA via a text-to-text transformer framework, and tracks both categorical slots and non-categorical slots in DST. In addition, we introduce two effective ways to construct unanswerable questions, namely, negative question sampling and context truncation, which enable our model to handle none value slots in the zero-shot DST setting. The extensive experiments show that our approaches substantially improve the existing zero-shot and few-shot results on MultiWoz. Moreover, compared to the fully trained baseline on the Schema-Guided Dialogue dataset, our approach shows better generalization ability in unseen domains.

Uncertainty Measures in Neural Belief Tracking and the Effects on Dialogue Policy Performance

Carel van Niekerk et al.

The ability to identify and resolve uncertainty is crucial for the robustness of a dialogue system. Indeed, this has been confirmed empirically on systems that utilise Bayesian approaches to dialogue belief tracking. However, such systems consider only confidence estimates and have difficulty scaling to more complex settings. Neural dialogue systems, on the other hand, rarely take uncertainties into account. They are therefore overconfident in their decisions and less robust. Moreover, the performance of the tracking task is often evaluated in isolation, without consideration of its effect on the downstream policy optimisation. We propose the use of different uncertainty measures in neural belief tracking. The effects of these measures on the downstream task of policy optimisation are evaluated by adding selected measures of uncertainty to the feature space of the policy and training policies through interaction with a user simulator. Both human and simulated user results show that incorporating these measures leads to improvements both of the performance and of the robustness of the downstream dialogue policy. This highlights the importance of developing neural dialogue belief trackers that take uncertainty into account.

Dynamic Forecasting of Conversation Derailment

Yova Kementchedjhieva and Anders Søgaard

Online conversations can sometimes take a turn for the worse, either due to systematic cultural differences, accidental misunderstandings, or mere malice. Automatically forecasting derailment in public online conversations provides an opportunity to take early action to moderate it. Previous work in this space is limited, and we extend it in several ways. We apply a pretrained language encoder to the task, which outperforms earlier approaches. We further experiment with shifting the training paradigm for the task from a static to a dynamic one to increase the forecast horizon. This approach shows mixed results: in a high-quality data setting, a longer average forecast horizon can be achieved at the cost of a small drop in F1; in a low-quality data setting, however, dynamic training propagates the noise and is highly detrimental to performance.

A Semantic Filter Based on Relations for Knowledge Graph Completion

Zongwei Liang et al.

Knowledge graph embedding, representing entities and relations in the knowledge graphs with high-dimensional vectors, has made significant progress in link prediction. More researchers have explored the representational capabilities of models in recent years. That is, they investigate better representational models to fit symmetry/antisymmetry and combination relationships. The current embedding models are more inclined to utilize the identical vector for the same entity in various triples to measure the matching performance. The observation that measuring the rationality of specific triples means comparing the matching degree of the specific attributes associated with the relations is well-known. Inspired by this fact, this paper designs Semantic Filter Based on Relations(SFBR) to extract the required attributes of the entities. Then the rationality of triples is compared under these extracted attributes through the traditional embedding models. The semantic filter module can be added to most geometric and tensor decomposition models with minimal additional memory. experiments on the benchmark datasets show that the semantic filter based on relations can suppress the impact of other attribute dimensions and improve link prediction performance. The tensor decomposition models with SFBR have achieved state-of-the-art.

AdapterDrop: On the Efficiency of Adapters in Transformers

Andreas Rücklé et al.

Transformer models are expensive to fine-tune, slow for inference, and have large storage requirements. Recent approaches tackle these shortcomings by training smaller models, dynamically reducing the model size, and by training light-weight adapters. In this paper, we propose AdapterDrop, removing adapters from lower transformer layers during training and inference, which incorporates concepts from all three directions. We show that AdapterDrop can dynamically reduce the computational overhead when performing inference over multiple tasks simultaneously, with minimal decrease in task performances. We further prune adapters from AdapterFusion, which improves the inference efficiency while maintaining the task performances entirely.

Understanding and Overcoming the Challenges of Efficient Transformer Quantization

Yelysei Bondarenko, Markus Nagel, and tijmen Blankevoort

Transformer-based architectures have become the de-facto standard models for a wide range of Natural Language Processing tasks. However, their memory footprint and high latency are prohibitive for efficient deployment and inference on resource-limited devices. In this work, we explore quantization for transformers. We show that transformers have unique quantization challenges – namely, high dynamic activation ranges that are difficult to represent with a low bit fixed-point format. We establish that these activations contain structured outliers in the residual connections that encourage specific attention patterns, such as attending to the special

separator token. To combat these challenges, we present three solutions based on post-training quantization and quantization-aware training, each with a different set of compromises for accuracy, model size, and ease of use. In particular, we introduce a novel quantization scheme – per-embedding-group quantization. We demonstrate the effectiveness of our methods on the GLUE benchmark using BERT, establishing state-of-the-art results for post-training quantization. Finally, we show that transformer weights and embeddings can be quantized to ultra-low bit-widths, leading to significant memory savings with a minimum accuracy loss. Our source code is available at `https://github.com/qualcomm-ai-research/transformer-quantization`.

CAPE: Context-Aware Private Embeddings for Private Language Learning
Richard Plant, Dimitra Gkatzia, and Valerio Giuffrida

Neural language models have contributed to state-of-the-art results in a number of downstream applications including sentiment analysis, intent classification and others. However, obtaining text representations or embeddings using these models risks encoding personally identifiable information learned from language and context cues that may lead to privacy leaks. To ameliorate this issue, we propose Context-Aware Private Embeddings (CAPE), a novel approach which combines differential privacy and adversarial learning to preserve privacy during training of embeddings. Specifically, CAPE firstly applies calibrated noise through differential privacy to maintain the privacy of text representations by preserving the encoded semantic links while obscuring sensitive information. Next, CAPE employs an adversarial training regime that obscures identified private variables. Experimental results demonstrate that our proposed approach is more effective in reducing private information leakage than either single intervention, with approximately a 3% reduction in attacker performance compared to the best-performing current method.

Text Detoxification using Large Pre-trained Neural Models
David Dale et al.

We present two novel unsupervised methods for eliminating toxicity in text. Our first method combines two recent ideas: (1) guidance of the generation process with small style-conditional language models and (2) use of paraphrasing models to perform style transfer. We use a well-performing paraphraser guided by style-trained language models to keep the text content and remove toxicity. Our second method uses BERT to replace toxic words with their non-offensive synonyms. We make the method more flexible by enabling BERT to replace mask tokens with a variable number of words. Finally, we present the first large-scale comparative study of style transfer models on the task of toxicity removal. We compare our models with a number of methods for style transfer. The models are evaluated in a reference-free way using a combination of unsupervised style transfer metrics. Both methods we suggest yield new SOTA results.

Document-Level Text Simplification: Dataset, Criteria and Baseline
Renliang Sun, Hanqi Jin, and Xiaojun Wan

Text simplification is a valuable technique. However, current research is limited to sentence simplification. In this paper, we define and investigate a new task of document-level text simplification, which aims to simplify a document consisting of multiple sentences. Based on Wikipedia dumps, we first construct a large-scale dataset named D-Wikipedia and perform analysis and human evaluation on it to show that the dataset is reliable. Then, we propose a new automatic evaluation metric called D-SARI that is more suitable for the document-level simplification task. Finally, we select several representative models as baseline models for this task and perform automatic evaluation and human evaluation. We analyze the results and point out the shortcomings of the baseline models.

A Bag of Tricks for Dialogue Summarization
Muhammad Khalifa, Miguel Ballesteros, and Kathleen McKeown

Dialogue summarization comes with its own peculiar challenges as opposed to news or scientific articles summarization. In this work, we explore four different challenges of the task: handling and differentiating parts of the dialogue belonging to multiple speakers, negation understanding, reasoning about the situation, and informal language understanding. Using a pretrained sequence-to-sequence language model, we explore speaker name substitution, negation scope highlighting, multi-task learning with relevant tasks, and pretraining on in-domain data. Our experiments show that our proposed techniques indeed improve summarization performance, outperforming strong baselines.

Paraphrasing Compound Nominalizations
John Lee, Ho Hung Lim, and Carol Webster

A nominalization uses a deverbal noun to describe an event associated with its underlying verb. Commonly found in academic and formal texts, nominalizations can be difficult to interpret because of ambiguous semantic relations between the deverbal noun and its arguments. Our goal is to interpret nominalizations by generating clausal paraphrases. We address compound nominalizations with both nominal and adjectival modifiers, as well as prepositional phrases. In evaluations on a number of unsupervised methods, we obtained the strongest performance by using a pre-trained contextualized language model to re-rank paraphrase candidates identified by a textual entailment model.

Data-QuestEval: A Referenceless Metric for Data-to-Text Semantic Evaluation
Clement Rebuffel et al.

QuestEval is a reference-less metric used in text-to-text tasks, that compares the generated summaries directly to the source text, by automatically asking and answering questions. Its adaptation to Data-to-Text tasks is not straightforward, as it requires multimodal Question Generation and Answering systems on the considered tasks, which are seldom available. To this purpose, we propose a method to build synthetic multimodal corpora enabling to train multimodal components for a data-QuestEval metric. The resulting metric is reference-less and multimodal; it obtains state-of-the-art correlations with human judgment on the WebNLG and WikiBio benchmarks. We make data-QuestEval's code and models available for reproducibility purpose, as part of the QuestEval project.

Low-Rank Subspaces for Unsupervised Entity Linking

Akhil Arora, Alberto Garcia-Duran, and Robert West

Entity linking is an important problem with many applications. Most previous solutions were designed for settings where annotated training data is available, which is, however, not the case in numerous domains. We propose a light-weight and scalable entity linking method, Eigenthemes, that relies solely on the availability of entity names and a referent knowledge base. Eigenthemes exploits the fact that the entities that are truly mentioned in a document (the "gold entities") tend to form a semantically dense subset of the set of all candidate entities in the document. Geometrically speaking, when representing entities as vectors via some given embedding, the gold entities tend to lie in a low-rank subspace of the full embedding space. Eigenthemes identifies this subspace using the singular value decomposition and scores candidate entities according to their proximity to the subspace. On the empirical front, we introduce multiple strong baselines that compare favorably to (and sometimes even outperform) the existing state of the art. Extensive experiments on benchmark datasets from a variety of real-world domains showcase the effectiveness of our approach.

TDEER: An Efficient Translating Decoding Schema for Joint Extraction of Entities and Relations

Xianming Li et al.

Joint extraction of entities and relations from unstructured texts to form factual triples is a fundamental task of constructing a Knowledge Base (KB). A common method is to decode triples by predicting entity pairs to obtain the corresponding relation. However, it is still challenging to handle this task efficiently, especially for the overlapping triple problem. To address such a problem, this paper proposes a novel efficient entities and relations extraction model called **TDEER**, which stands for **T**ranslating **D**ecoding Schema for Joint **E**xtraction of **E**ntities and **R**elations. Unlike the common approaches, the proposed translating decoding schema regards the relation as a translating operation from subject to objects, i.e., TDEER decodes triples as $subject + relation \rightarrow objects$. TDEER can naturally handle the overlapping triple problem, because the translating decoding schema can recognize all possible triples, including overlapping and non-overlapping triples. To enhance model robustness, we introduce negative samples to alleviate error accumulation at different stages. Extensive experiments on public datasets demonstrate that TDEER produces competitive results compared with the state-of-the-art (SOTA) baselines. Furthermore, the computation complexity analysis indicates that TDEER is more efficient than powerful baselines. Especially, the proposed TDEER is 2 times faster than the recent SOTA models. The code is available at `https://github.com/4AI/TDEER`.

Extracting Event Temporal Relations via Hyperbolic Geometry

Xingwei Tan, Gabriele Pergola, and Yulan He

Detecting events and their evolution through time is a crucial task in natural language understanding. Recent neural approaches to event temporal relation extraction typically map events to embeddings in the Euclidean space and train a classifier to detect temporal relations between event pairs. However, embeddings in the Euclidean space cannot capture richer asymmetric relations such as event temporal relations. We thus propose to embed events into hyperbolic spaces, which are intrinsically oriented at modeling hierarchical structures. We introduce two approaches to encode events and their temporal relations in hyperbolic spaces. One approach leverages hyperbolic embeddings to directly infer event relations through simple geometrical operations. In the second one, we devise an end-to-end architecture composed of hyperbolic neural units tailored for the temporal relation extraction task. Thorough experimental assessments on widely used datasets have shown the benefits of revisiting the tasks on a different geometrical space, resulting in state-of-the-art performance on several standard metrics. Finally, the ablation study and several qualitative analyses highlighted the rich event semantics implicitly encoded into hyperbolic spaces.

Honey or Poison? Solving the Trigger Curse in Few-shot Event Detection via Causal Intervention

Jiawei Chen et al.

Event detection has long been troubled by the trigger curse: overfitting the trigger will harm the generalization ability while underfitting it will hurt the detection performance. This problem is even more severe in few-shot scenario. In this paper, we identify and solve the trigger curse problem in few-shot event detection (FSED) from a causal view. By formulating FSED with a structural causal model (SCM), we found that the trigger is a confounder of the context and the result, which makes previous FSED methods much easier to overfit triggers. To resolve this problem, we propose to intervene on the context via backdoor adjustment during training. Experiments show that our method significantly improves the FSED on both ACE05 and MAVEN datasets.

Matching-oriented Embedding Quantization For Ad-hoc Retrieval
Shitao Xiao et al.

Product quantization (PQ) is a widely used technique for ad-hoc retrieval. Recent studies propose supervised PQ, where the embedding and quantization models can be jointly trained with supervised learning. However, there is a lack of appropriate formulation of the joint training objective; thus, the improvements over previous non-supervised baselines are limited in reality. In this work, we propose the Matching-oriented Product Quantization (MoPQ), where a novel objective Multinoulli Contrastive Loss (MCL) is formulated. With the minimization of MCL, we are able to maximize the matching probability of query and ground-truth key, which contributes to the optimal retrieval accuracy. Given that the exact computation of MCL is intractable due to the demand of vast contrastive samples, we further propose the Differentiable Cross-device Sampling (DCS), which significantly augments the contrastive samples for precise approximation of MCL. We conduct extensive experimental studies on four real-world datasets, whose results verify the effectiveness of MoPQ. The code is available at https://github.com/microsoft/MoPQ.

Efficient Mind-Map Generation via Sequence-to-Graph and Reinforced Graph Refinement
Mengting Hu et al.

A mind-map is a diagram that represents the central concept and key ideas in a hierarchical way. Converting plain text into a mind-map will reveal its key semantic structure and be easier to understand. Given a document, the existing automatic mind-map generation method extracts the relationships of every sentence pair to generate the directed semantic graph for this document. The computation complexity increases exponentially with the length of the document. Moreover, it is difficult to capture the overall semantics. To deal with the above challenges, we propose an efficient mind-map generation network that converts a document into a graph via sequence-to-graph. To guarantee a meaningful mind-map, we design a graph refinement module to adjust the relation graph in a reinforcement learning manner. Extensive experimental results demonstrate that the proposed approach is more effective and efficient than the existing methods. The inference time is reduced by thousands of times compared with the existing methods. The case studies verify that the generated mind-maps better reveal the underlying semantic structures of the document.

Deep Attention Diffusion Graph Neural Networks for Text Classification
Yonghao Liu et al.

Text classification is a fundamental task with broad applications in natural language processing. Recently, graph neural networks (GNNs) have attracted much attention due to their powerful representation ability. However, most existing methods for text classification based on GNNs consider only one-hop neighborhoods and low-frequency information within texts, which cannot fully utilize the rich context information of documents. Moreover, these models suffer from over-smoothing issues if many graph layers are stacked. In this paper, a Deep Attention Diffusion Graph Neural Network (DADGNN) model is proposed to learn text representations, bridging the chasm of interaction difficulties between a word and its distant neighbors. Experimental results on various standard benchmark datasets demonstrate the superior performance of the present approach.

Balancing Methods for Multi-label Text Classification with Long-Tailed Class Distribution
Yi Huang et al.

Multi-label text classification is a challenging task because it requires capturing label dependencies. It becomes even more challenging when class distribution is long-tailed. Resampling and re-weighting are common approaches used for addressing the class imbalance problem, however, they are not effective when there is label dependency besides class imbalance because they result in oversampling of common labels. Here, we introduce the application of balancing loss functions for multi-label text classification. We perform experiments on a general domain dataset with 90 labels (Reuters-21578) and a domain-specific dataset from PubMed with 18211 labels. We find that a distribution-balanced loss function, which inherently addresses both the class imbalance and label linkage problems, outperforms commonly used loss functions. Distribution balancing methods have been successfully used in the image recognition field. Here, we show their effectiveness in natural language processing. Source code is available at https://github.com/blessu/BalancedLossNLP.

Bayesian Topic Regression for Causal Inference
Maximilian Ahrens et al.

Causal inference using observational text data is becoming increasingly popular in many research areas. This paper presents the Bayesian Topic Regression (BTR) model that uses both text and numerical information to model an outcome variable. It allows estimation of both discrete and continuous treatment effects. Furthermore, it allows for the inclusion of additional numerical confounding factors next to text data. To this end, we combine a supervised Bayesian topic model with a Bayesian regression framework and perform supervised representation learning for the text features jointly with the regression parameter training, respecting the Frisch-Waugh-Lovell theorem. Our paper makes two main contributions. First, we provide a regression framework that allows causal inference in settings when both text and numerical confounders are of relevance. We show with synthetic and semi-synthetic datasets that our joint approach recovers ground truth with lower bias than any benchmark model, when text and numerical features are correlated. Second, experiments on two real-world datasets demonstrate that a joint and supervised learning strategy also yields superior prediction results compared to strategies that estimate regression weights for text and non-text features separately, being even competitive with more complex deep neural networks.

Enjoy the Salience: Towards Better Transformer-based Faithful Explanations with Word Salience
George Chrysostomou and Nikolaos Aletras

Pretrained transformer-based models such as BERT have demonstrated state-of-the-art predictive performance when adapted into a range of natural language processing tasks. An open problem is how to improve the faithfulness of explanations (rationales) for the predictions of these models. In this paper, we hypothesize that salient information extracted a priori from the training data can complement the task-specific information learned by the model during fine-tuning on a downstream task. In this way, we aim to help BERT not to forget assigning importance to informative input tokens when making predictions by proposing SaLoss; an auxiliary loss function for guiding the multi-head attention mechanism during training to be close to salient information extracted a priori using TextRank. Experiments for explanation faithfulness across five datasets, show that models trained with SaLoss consistently provide more faithful explanations across four different feature attribution methods compared to vanilla BERT. Using the rationales extracted from vanilla BERT and SaLoss models to train inherently faithful classifiers, we further show that the latter result in higher predictive performance in downstream tasks.

What's in Your Head? Emergent Behaviour in Multi-Task Transformer Models
Mor Geva et al.

The primary paradigm for multi-task training in natural language processing is to represent the input with a shared pre-trained language model, and add a small, thin network (head) per task. Given an input, a target head is the head that is selected for outputting the final prediction. In this work, we examine the behaviour of non-target heads, that is, the output of heads when given input that belongs to a different task than the one they were trained for. We find that non-target heads exhibit emergent behaviour, which may either explain the target task, or generalize beyond their original task. For example, in a numerical reasoning task, a span extraction head extracts from the input the arguments to a computation that results in a number generated by a target generative head. In addition, a summarization head that is trained with a target question answering head, outputs query-based summaries when given a question and a context from which the answer is to be extracted. This emergent behaviour suggests that multi-task training leads to non-trivial extrapolation of skills, which can be harnessed for interpretability and generalization.

Don't Search for a Search Method — Simple Heuristics Suffice for Adversarial Text Attacks
Nathaniel Berger et al.

Recently more attention has been given to adversarial attacks on neural networks for natural language processing (NLP). A central research topic has been the investigation of search algorithms and search constraints, accompanied by benchmark algorithms and tasks. We implement an algorithm inspired by zeroth order optimization-based attacks and compare with the benchmark results in the TextAttack framework. Surprisingly, we find that optimization-based methods do not yield any improvement in a constrained setup and slightly benefit from approximate gradient information only in unconstrained setups where search spaces are larger. In contrast, simple heuristics exploiting nearest neighbors without querying the target function yield substantial success rates in constrained setups, and nearly full success rate in unconstrained setups, at an order of magnitude fewer queries. We conclude from these results that current TextAttack benchmark tasks are too easy and constraints are too strict, preventing meaningful research on black-box adversarial text attacks.

Adversarial Attacks on Knowledge Graph Embeddings via Instance Attribution Methods
Peru Bhardwaj et al.

Despite the widespread use of Knowledge Graph Embeddings (KGE), little is known about the security vulnerabilities that might disrupt their intended behaviour. We study data poisoning attacks against KGE models for link prediction. These attacks craft adversarial additions or deletions at training time to cause model failure at test time. To select adversarial deletions, we propose to use the model-agnostic instance attribution methods from Interpretable Machine Learning, which identify the training instances that are most influential to a neural model's predictions on test instances. We use these influential triples as adversarial deletions. We further propose a heuristic method to replace one of the two entities in each influential triple to generate adversarial additions. Our experiments show that the proposed strategies outperform the state-of-art data poisoning attacks on KGE models and improve the MRR degradation due to the attacks by up to 62% over the baselines.

Locke's Holiday: Belief Bias in Machine Reading
Anders Søgaard

I highlight a simple failure mode of state-of-the-art machine reading systems: when contexts do not align with commonly shared beliefs. For example, machine reading systems fail to answer *What did Elizabeth want?* correctly in the context of 'My kingdom for a cough drop, cried Queen Elizabeth.' Biased by co-occurrence statistics in the training data of pretrained language models, systems predict *my kingdom*, rather than *a cough drop*. I argue such biases are analogous to human belief biases and present a carefully designed challenge dataset for English machine reading, called AUTO-LOCKE, to quantify such effects. Evaluations of machine reading systems on AUTO-LOCKE show the pervasiveness of belief bias in machine reading.

Sequence Length is a Domain: Length-based Overfitting in Transformer Models
Dusan Varis and Ondřej Bojar

Transformer-based sequence-to-sequence architectures, while achieving state-of-the-art results on a large number of NLP tasks, can still suffer from overfitting during training. In practice, this is usually countered either by applying regularization methods (e.g. dropout, L2-regularization) or by providing huge amounts of training data. Additionally, Transformer and other architectures are known to struggle when generating very long sequences. For example, in machine translation, the neural-based systems perform worse on very long sequences when compared to the preceding phrase-based translation approaches (Koehn and Knowles, 2017). We present results which suggest that the issue might also be in the mismatch between the length distributions of the training and validation data combined with the aforementioned tendency of the neural networks to overfit to the training data. We demonstrate on a simple string editing tasks and a machine translation task that the Transformer model performance drops significantly when facing sequences of length diverging from the length distribution in the training data. Additionally, we show that the observed drop in performance is due to the hypothesis length corresponding to the lengths seen by the model during training rather than the length of the input sequence.

Contrasting Human- and Machine-Generated Word-Level Adversarial Examples for Text Classification

Maximilian Mozes et al.

Research shows that natural language processing models are generally considered to be vulnerable to adversarial attacks; but recent work has drawn attention to the issue of validating these adversarial inputs against certain criteria (e.g., the preservation of semantics and grammaticality). Enforcing constraints to uphold such criteria may render attacks unsuccessful, raising the question of whether valid attacks are actually feasible. In this work, we investigate this through the lens of human language ability. We report on crowdsourcing studies in which we task humans with iteratively modifying words in an input text, while receiving immediate model feedback, with the aim of causing a sentiment classification model to misclassify the example. Our findings suggest that humans are capable of generating a substantial amount of adversarial examples using semantics-preserving word substitutions. We analyze how human-generated adversarial examples compare to the recently proposed TextFooler, Genetic, BAE and SememePSO attack algorithms on the dimensions naturalness, preservation of sentiment, grammaticality and substitution rate. Our findings suggest that human-generated adversarial examples are not more able than the best algorithms to generate natural-reading, sentiment-preserving examples, though they do so by being much more computationally efficient.

Is Information Density Uniform in Task-Oriented Dialogues?

Mario Giulianelli, Arabella Sinclair, and Raquel Fernández

The Uniform Information Density principle states that speakers plan their utterances to reduce fluctuations in the density of the information transmitted. In this paper, we test whether, and within which contextual units this principle holds in task-oriented dialogues. We show that there is evidence supporting the principle in written dialogues where participants play a cooperative reference game as well as in spoken dialogues involving instruction giving and following. Our study underlines the importance of identifying the relevant contextual components, showing that information content increases particularly within topically and referentially related contextual units.

On Homophony and Rényi Entropy

Tiago Pimentel et al.

Homophony's widespread presence in natural languages is a controversial topic. Recent theories of language optimality have tried to justify its prevalence, despite its negative effects on cognitive processing time, e.g., Piantadosi et al. (2012) argued homophony enables the reuse of efficient wordforms and is thus beneficial for languages. This hypothesis has recently been challenged by Trott and Bergen (2020), who posit that good wordforms are more often homophonous simply because they are more phonotactically probable. In this paper, we join in on the debate. We first propose a new information-theoretic quantification of a language's homophony: the sample Rényi entropy. Then, we use this quantification to revisit Trott and Bergen's claims. While their point is theoretically sound, a specific methodological issue in their experiments raises doubts about their results. After addressing this issue, we find no clear pressure either towards or against homophony—a much more nuanced result than either Piantadosi et al.'s or Trott and Bergen's findings.

Synthetic Textual Features for the Large-Scale Detection of Basic-level Categories in English and Mandarin

Yiwen Chen and Simone Teufel

Basic-level categories (BLC) are an important psycholinguistic concept introduced by Rosch et al. (1976); they are defined as the most inclusive categories for which a concrete mental image of the category as a whole can be formed, and also as those categories which are acquired early in life. Rosch's original algorithm for detecting BLC (called cue-validity) is based on the availability of semantic features such as "has tail" for "cat", and has remained untested at large. An at-scale algorithm for the automatic determination of BLC exists, but it operates without Rosch-style semantic features, and is thus unable to verify Rosch's hypothesis. We present the first method for the detection of BLC at scale that makes use of Rosch-style semantic features. For both English and Mandarin, we test three methods of generating such features for any synset within Wordnet (WN): extraction of textual features from Wikipedia pages, Distributional Memory (DM) and BART. The best of our methods outperforms the current SoA in BLC detection, with an accuracy of English BLC detection of 75.0%, and of Mandarin BLC detection 80.7% on a test set. When applied to all of WordNet, our model predicts that 1,118 synsets in English Wordnet (1.4%) are BLC, far fewer than existing methods, and with a precision

improvement of over 200% over these. As well as confirming the usefulness of Rosch's cue validity algorithm, we also developed and evaluated our own new indicator for BLC, which models the fact that BLC features tend to be BLC themselves.

TimeTraveler: Reinforcement Learning for Temporal Knowledge Graph Forecasting

Haohai Sun et al.

Temporal knowledge graph (TKG) reasoning is a crucial task that has gained increasing research interest in recent years. Most existing methods focus on reasoning at past timestamps to complete the missing facts, and there are only a few works of reasoning on known TKGs to forecast future facts. Compared with the completion task, the forecasting task is more difficult that faces two main challenges: (1) how to effectively model the time information to handle future timestamps? (2) how to make inductive inference to handle previously unseen entities that emerge over time? To address these challenges, we propose the first reinforcement learning method for forecasting. Specifically, the agent travels on historical knowledge graph snapshots to search for the answer. Our method defines a relative time encoding function to capture the timespan information, and we design a novel time-shaped reward based on Dirichlet distribution to guide the model learning. Furthermore, we propose a novel representation method for unseen entities to improve the inductive inference ability of the model. We evaluate our method for this link prediction task at future timestamps. Extensive experiments on four benchmark datasets demonstrate substantial performance improvement meanwhile with higher explainability, less calculation, and fewer parameters when compared with existing state-of-the-art methods.

Code-switched inspired losses for spoken dialog representations

Pierre Colombo et al.

Spoken dialogue systems need to be able to handle both multiple languages and multilinguality inside a conversation (*e.g* in case of code-switching). In this work, we introduce new pretraining losses tailored to learn generic multilingual spoken dialogue representations. The goal of these losses is to expose the model to code-switched language. In order to scale up training, we automatically build a pretraining corpus composed of multilingual conversations in five different languages (French, Italian, English, German and Spanish) from `OpenSubtitles`, a huge multilingual corpus composed of 24.3G tokens. We test the generic representations on `MIAM`, a new benchmark composed of five dialogue act corpora on the same aforementioned languages as well as on two novel multilingual tasks (*i.e* multilingual mask utterance retrieval and multilingual inconsistency identification). Our experiments show that our new losses achieve a better performance in both monolingual and multilingual settings.

BiQUE: Biquaternionic Embeddings of Knowledge Graphs

Jia Guo and Stanley Kok

Knowledge graph embeddings (KGEs) compactly encode multi-relational knowledge graphs (KGs). Existing KGE models rely on geometric operations to model relational patterns. Euclidean (circular) rotation is useful for modeling patterns such as symmetry, but cannot represent hierarchical semantics. In contrast, hyperbolic models are effective at modeling hierarchical relations, but do not perform as well on patterns on which circular rotation excels. It is crucial for KGE models to unify multiple geometric transformations so as to fully cover the multifarious relations in KGs. To do so, we propose BiQUE, a novel model that employs *biquaternions* to integrate multiple geometric transformations, viz., scaling, translation, Euclidean rotation, and hyperbolic rotation. BiQUE makes the best trade-offs among geometric operators during training, picking the best one (or their best combination) for each relation. Experiments on five datasets show BiQUE's effectiveness.

Learning Neural Ordinary Equations for Forecasting Future Links on Temporal Knowledge Graphs

Zhen Han et al.

There has been an increasing interest in inferring future links on temporal knowledge graphs (KG). While links on temporal KGs vary continuously over time, the existing approaches model the temporal KGs in discrete state spaces. To this end, we propose a novel continuum model by extending the idea of neural ordinary differential equations (ODEs) to multi-relational graph convolutional networks. The proposed model preserves the continuous nature of dynamic multi-relational graph data and encodes both temporal and structural information into continuous-time dynamic embeddings. In addition, a novel graph transition layer is applied to capture the transitions on the dynamic graph, i.e., edge formation and dissolution. We perform extensive experiments on five benchmark datasets for temporal KG reasoning, showing our model's superior performance on the future link forecasting task.

RAP: Robustness-Aware Perturbations for Defending against Backdoor Attacks on NLP Models

Wenkai Yang et al.

Backdoor attacks, which maliciously control a well-trained model's outputs of the instances with specific triggers, are recently shown to be serious threats to the safety of reusing deep neural networks (DNNs). In this work, we propose an efficient online defense mechanism based on robustness-aware perturbations. Specifically, by analyzing the backdoor training process, we point out that there exists a big gap of robustness between poisoned and clean samples. Motivated by this observation, we construct a word-based robustness-aware perturbation to distinguish poisoned samples from clean samples to defend against the backdoor attacks on natural language processing (NLP) models. Moreover, we give a theoretical analysis about the feasibility of

our robustness-aware perturbation-based defense method. Experimental results on sentiment analysis and toxic detection tasks show that our method achieves better defending performance and much lower computational costs than existing online defense methods. Our code is available at https://github.com/lancopku/RAP.

A Strong Baseline for Query Efficient Attacks in a Black Box Setting
Rishabh Maheshwary, Saket Maheshwary, and Vikram Pudi

Existing black box search methods have achieved high success rate in generating adversarial attacks against NLP models. However, such search methods are inefficient as they do not consider the amount of queries required to generate adversarial attacks. Also, prior attacks do not maintain a consistent search space while comparing different search methods. In this paper, we propose a query efficient attack strategy to generate plausible adversarial examples on text classification and entailment tasks. Our attack jointly leverages attention mechanism and locality sensitive hashing (LSH) to reduce the query count. We demonstrate the efficacy of our approach by comparing our attack with four baselines across three different search spaces. Further, we benchmark our results across the same search space used in prior attacks. In comparison to attacks proposed, on an average, we are able to reduce the query count by 75% across all datasets and target models. We also demonstrate that our attack achieves a higher success rate when compared to prior attacks in a limited query setting.

Machine Translation Decoding beyond Beam Search
Rémi Leblond et al.

Beam search is the go-to method for decoding auto-regressive machine translation models. While it yields consistent improvements in terms of BLEU, it is only concerned with finding outputs with high model likelihood, and is thus agnostic to whatever end metric or score practitioners care about. Our aim is to establish whether beam search can be replaced by a more powerful metric-driven search technique. To this end, we explore numerous decoding algorithms, including some which rely on a value function parameterised by a neural network, and report results on a variety of metrics. Notably, we introduce a Monte-Carlo Tree Search (MCTS) based method and showcase its competitiveness. We provide a blueprint for how to use MCTS fruitfully in language applications, which opens promising future directions. We find that which algorithm is best heavily depends on the characteristics of the goal metric; we believe that our extensive experiments and analysis will inform further research in this area.

Document Graph for Neural Machine Translation
Mingzhou Xu et al.

Previous works have shown that contextual information can improve the performance of neural machine translation (NMT). However, most existing document-level NMT methods failed to leverage contexts beyond a few set of previous sentences. How to make use of the whole document as global contexts is still a challenge. To address this issue, we hypothesize that a document can be represented as a graph that connects relevant contexts regardless of their distances. We employ several types of relations, including adjacency, syntactic dependency, lexical consistency, and coreference, to construct the document graph. Then, we incorporate both source and target graphs into the conventional Transformer architecture with graph convolutional networks. Experiments on various NMT benchmarks, including IWSLT English–French, Chinese-English, WMT English–German and Opensubtitle English–Russian, demonstrate that using document graphs can significantly improve the translation quality. Extensive analysis verifies that the document graph is beneficial for capturing discourse phenomena.

An Empirical Investigation of Word Alignment Supervision for Zero-Shot Multilingual Neural Machine Translation
Alessandro Raganato et al.

Zero-shot translations is a fascinating feature of Multilingual Neural Machine Translation (MNMT) systems. These MNMT models are usually trained on English-centric data, i.e. English either as the source or target language, and with a language label prepended to the input indicating the target language. However, recent work has highlighted several flaws of these models in zero-shot scenarios where language labels are ignored and the wrong language is generated or different runs show highly unstable results. In this paper, we investigate the benefits of an explicit alignment to language labels in Transformer-based MNMT models in the zero-shot context, by jointly training one cross attention head with word alignment supervision to stress the focus on the target language label. We compare and evaluate several MNMT systems on three multilingual MT benchmarks of different sizes, showing that simply supervising one cross attention head to focus both on word alignments and language labels reduces the bias towards translating into the wrong language, improving the zero-shot performance overall. Moreover, as an additional advantage, we find that our alignment supervision leads to more stable results across different training runs.

Graph Algorithms for Multiparallel Word Alignment
Ayyoob ImaniGooghari et al.

With the advent of end-to-end deep learning approaches in machine translation, interest in word alignments initially decreased; however, they have again become a focus of research more recently. Alignments are useful for typological research, transferring formatting like markup to translated texts, and can be used in the decoding of machine translation systems. At the same time, massively multilingual processing is becoming an important NLP scenario, and pretrained language and machine translation models that are truly multilingual

are proposed. However, most alignment algorithms rely on bitexts only and do not leverage the fact that many parallel corpora are multiparallel. In this work, we exploit the multiparallelity of corpora by representing an initial set of bilingual alignments as a graph and then predicting additional edges in the graph. We present two graph algorithms for edge prediction: one inspired by recommender systems and one based on network link prediction. Our experimental results show absolute improvements in F1 of up to 28% over the baseline bilingual word aligner in different datasets.

Improving the Quality Trade-Off for Neural Machine Translation Multi-Domain Adaptation
Eva Hasler et al.

Building neural machine translation systems to perform well on a specific target domain is a well-studied problem. Optimizing system performance for multiple, diverse target domains however remains a challenge. We study this problem in an adaptation setting where the goal is to preserve the existing system quality while incorporating data for domains that were not the focus of the original translation system. We find that we can improve over the performance trade-off offered by Elastic Weight Consolidation with a relatively simple data mixing strategy. At comparable performance on the new domains, catastrophic forgetting is mitigated significantly on strong WMT baselines. Combining both approaches improves the Pareto frontier on this task.

Language Modeling, Lexical Translation, Reordering: The Training Process of NMT through the Lens of Classical SMT
Elena Voita, Rico Sennrich, and Ivan Titov

Differently from the traditional statistical MT that decomposes the translation task into distinct separately learned components, neural machine translation uses a single neural network to model the entire translation process. Despite neural machine translation being de-facto standard, it is still not clear how NMT models acquire different competences over the course of training, and how this mirrors the different models in traditional SMT. In this work, we look at the competences related to three core SMT components and find that during training, NMT first focuses on learning target-side language modeling, then improves translation quality approaching word-by-word translation, and finally learns more complicated reordering patterns. We show that this behavior holds for several models and language pairs. Additionally, we explain how such an understanding of the training process can be useful in practice and, as an example, show how it can be used to improve vanilla non-autoregressive neural machine translation by guiding teacher model selection.

Effective Fine-Tuning Methods for Cross-lingual Adaptation
Tao Yu and Shafiq Joty

Large scale multilingual pre-trained language models have shown promising results in zero- and few-shot cross-lingual tasks. However, recent studies have shown their lack of generalizability when the languages are structurally dissimilar. In this work, we propose a novel fine-tuning method based on co-training that aims to learn more generalized semantic equivalences as a complementary to multilingual language modeling using the unlabeled data in the target language. We also propose an adaption method based on contrastive learning to better capture the semantic relationship in the parallel data, when a few translation pairs are available. To show our method's effectiveness, we conduct extensive experiments on cross-lingual inference and review classification tasks across various languages. We report significant gains compared to directly fine-tuning multilingual pre-trained models and other semi-supervised alternatives.

Rethinking Data Augmentation for Low-Resource Neural Machine Translation: A Multi-Task Learning Approach
Víctor M. Sánchez-Cartagena et al.

In the context of neural machine translation, data augmentation (DA) techniques may be used for generating additional training samples when the available parallel data are scarce. Many DA approaches aim at expanding the support of the empirical data distribution by generating new sentence pairs that contain infrequent words, thus making it closer to the true data distribution of parallel sentences. In this paper, we propose to follow a completely different approach and present a multi-task DA approach in which we generate new sentence pairs with transformations, such as reversing the order of the target sentence, which produce unfluent target sentences. During training, these augmented sentences are used as auxiliary tasks in a multi-task framework with the aim of providing new contexts where the target prefix is not informative enough to predict the next word. This strengthens the encoder and forces the decoder to pay more attention to the source representations of the encoder. Experiments carried out on six low-resource translation tasks show consistent improvements over the baseline and over DA methods aiming at extending the support of the empirical data distribution. The systems trained with our approach rely more on the source tokens, are more robust against domain shift and suffer less hallucinations.

Wino-X: Multilingual Winograd Schemas for Commonsense Reasoning and Coreference Resolution
Denis Emelin and Rico Sennrich

Winograd schemas are a well-established tool for evaluating coreference resolution (CoR) and commonsense reasoning (CSR) capabilities of computational models. So far, schemas remained largely confined to English, limiting their utility in multilingual settings. This work presents Wino-X, a parallel dataset of German, French, and Russian schemas, aligned with their English counterparts. We use this resource to investigate whether neu-

ral machine translation (NMT) models can perform CoR that requires commonsense knowledge and whether multilingual language models (MLLMs) are capable of CSR across multiple languages. Our findings show Wino-X to be exceptionally challenging for NMT systems that are prone to undesirable biases and unable to detect disambiguating information. We quantify biases using established statistical methods and define ways to address both of these issues. We furthermore present evidence of active cross-lingual knowledge transfer in MLLMs, whereby fine-tuning models on English schemas yields CSR improvements in other languages.

Discrete and Soft Prompting for Multilingual Models
Mengjie Zhao and Hinrich Schütze

It has been shown for English that discrete and soft prompting perform strongly in few-shot learning with pretrained language models (PLMs). In this paper, we show that discrete and soft prompting perform better than finetuning in multilingual cases: Crosslingual transfer and in-language training of multilingual natural language inference. For example, with 48 English training examples, finetuning obtains 33.74% accuracy in crosslingual transfer, barely surpassing the majority baseline (33.33%). In contrast, discrete and soft prompting outperform finetuning, achieving 36.43% and 38.79%. We also demonstrate good performance of prompting with training data in multiple languages other than English.

Vision Matters When It Should: Sanity Checking Multimodal Machine Translation Models
Jiaoda Li, Duygu Ataman, and Rico Sennrich

Multimodal machine translation (MMT) systems have been shown to outperform their text-only neural machine translation (NMT) counterparts when visual context is available. However, recent studies have also shown that the performance of MMT models is only marginally impacted when the associated image is replaced with an unrelated image or noise, which suggests that the visual context might not be exploited by the model at all. We hypothesize that this might be caused by the nature of the commonly used evaluation benchmark, also known as Multi30K, where the translations of image captions were prepared without actually showing the images to human translators. In this paper, we present a qualitative study that examines the role of datasets in stimulating the leverage of visual modality and we propose methods to highlight the importance of visual signals in the datasets which demonstrate improvements in reliance of models on the source images. Our findings suggest the research on effective MMT architectures is currently impaired by the lack of suitable datasets and careful consideration must be taken in creation of future MMT datasets, for which we also provide useful insights.

Efficient Inference for Multilingual Neural Machine Translation
Alexandre Berard et al.

Multilingual NMT has become an attractive solution for MT deployment in production. But to match bilingual quality, it comes at the cost of larger and slower models. In this work, we consider several ways to make multilingual NMT faster at inference without degrading its quality. We experiment with several "light decoder" architectures in two 20-language multi-parallel settings: small-scale on TED Talks and large-scale on ParaCrawl. Our experiments demonstrate that combining a shallow decoder with vocabulary filtering leads to almost 2 times faster inference with no loss in translation quality. We validate our findings with BLEU and chrF (on 380 language pairs), robustness evaluation and human evaluation.

Role of Language Relatedness in Multilingual Fine-tuning of Language Models: A Case Study in Indo-Aryan Languages
Tejas Dhamecha et al.

We explore the impact of leveraging the relatedness of languages that belong to the same family in NLP models using multilingual fine-tuning. We hypothesize and validate that multilingual fine-tuning of pre-trained language models can yield better performance on downstream NLP applications, compared to models fine-tuned on individual languages. A first of its kind detailed study is presented to track performance change as languages are added to a base language in a graded and greedy (in the sense of best boost of performance) manner; which reveals that careful selection of subset of related languages can significantly improve performance than utilizing all related languages. The Indo-Aryan (IA) language family is chosen for the study, the exact languages being Bengali, Gujarati, Hindi, Marathi, Oriya, Punjabi and Urdu. The script barrier is crossed by simple rule-based transliteration of the text of all languages to Devanagari. Experiments are performed on mBERT, IndicBERT, MuRIL and two RoBERTa-based LMs, the last two being pre-trained by us. Low resource languages, such as Oriya and Punjabi, are found to be the largest beneficiaries of multilingual fine-tuning. Textual Entailment, Entity Classification, Section Title Prediction, tasks of IndicGLUE and POS tagging form our test bed. Compared to monolingual fine tuning we get relative performance improvement of up to 150% in the downstream tasks. The surprise take-away is that for any language there is a particular combination of other languages which yields the best performance, and any additional language is in fact detrimental.

Comparing Feature-Engineering and Feature-Learning Approaches for Multilingual Translationese Classification
Daria Pylypenko et al.

Traditional hand-crafted linguistically-informed features have often been used for distinguishing between translated and original non-translated texts. By contrast, to date, neural architectures without manual feature engineering have been less explored for this task. In this work, we (i) compare the traditional feature-engineering-based approach to the feature-learning-based one and (ii) analyse the neural architectures in order

to investigate how well the hand-crafted features explain the variance in the neural models' predictions. We use pre-trained neural word embeddings, as well as several end-to-end neural architectures in both monolingual and multilingual settings and compare them to feature-engineering-based SVM classifiers. We show that (i) neural architectures outperform other approaches by more than 20 accuracy points, with the BERT-based model performing the best in both the monolingual and multilingual settings; (ii) while many individual hand-crafted translationese features correlate with neural model predictions, feature importance analysis shows that the most important features for neural and classical architectures differ; and (iii) our multilingual experiments provide empirical evidence for translationese universals across languages.

Multi-Sentence Resampling: A Simple Approach to Alleviate Dataset Length Bias and Beam-Search Degradation
Ivan Provilkov and Andrey Malinin

Neural Machine Translation (NMT) is known to suffer from a beam-search problem: after a certain point, increasing beam size causes an overall drop in translation quality. This effect is especially pronounced for long sentences. While much work was done analyzing this phenomenon, primarily for autoregressive NMT models, there is still no consensus on its underlying cause. In this work, we analyze errors that cause major quality degradation with large beams in NMT and Automatic Speech Recognition (ASR). We show that a factor that strongly contributes to the quality degradation with large beams is dataset length-bias - NMT datasets are strongly biased towards short sentences. To mitigate this issue, we propose a new data augmentation technique — Multi-Sentence Resampling (MSR). This technique extends the training examples by concatenating several sentences from the original dataset to make a long training example. We demonstrate that MSR significantly reduces degradation with growing beam size and improves final translation quality on the IWSTL15 En-Vi, IWSTL17 En-Fr, and WMT14 En-De datasets.

Cross-Policy Compliance Detection via Question Answering
Marzieh Saeidi, Majid Yazdani, and Andreas Vlachos

Policy compliance detection is the task of ensuring that a scenario conforms to a policy (e.g. a claim is valid according to government rules or a post in an online platform conforms to community guidelines). This task has been previously instantiated as a form of textual entailment, which results in poor accuracy due to the complexity of the policies. In this paper we propose to address policy compliance detection via decomposing it into question answering, where questions check whether the conditions stated in the policy apply to the scenario, and an expression tree combines the answers to obtain the label. Despite the initial upfront annotation cost, we demonstrate that this approach results in better accuracy, especially in the cross-policy setup where the policies during testing are unseen in training. In addition, it allows us to use existing question answering models pre-trained on existing large datasets. Finally, it explicitly identifies the information missing from a scenario in case policy compliance cannot be determined. We conduct our experiments using a recent dataset consisting of government policies, which we augment with expert annotations and find that the cost of annotating question answering decomposition is largely offset by improved inter-annotator agreement and speed.

Meta-LMTC: Meta-Learning for Large-Scale Multi-Label Text Classification
Ran Wang et al.

Large-scale multi-label text classification (LMTC) tasks often face long-tailed label distributions, where many labels have few or even no training instances. Although current methods can exploit prior knowledge to handle these few/zero-shot labels, they neglect the meta-knowledge contained in the dataset that can guide models to learn with few samples. In this paper, for the first time, this problem is addressed from a meta-learning perspective. However, the simple extension of meta-learning approaches to multi-label classification is sub-optimal for LMTC tasks due to long-tailed label distribution and coexisting of few- and zero-shot scenarios. We propose a meta-learning approach named META-LMTC. Specifically, it constructs more faithful and more diverse tasks according to well-designed sampling strategies and directly incorporates the objective of adapting to new low-resource tasks into the meta-learning phase. Extensive experiments show that META-LMTC achieves state-of-the-art performance against strong baselines and can still enhance powerful BERTlike models.

Unsupervised Multi-View Post-OCR Error Correction With Language Models
Harsh Gupta et al.

We investigate post-OCR correction in a setting where we have access to different OCR views of the same document. The goal of this study is to understand if a pretrained language model (LM) can be used in an unsupervised way to reconcile the different OCR views such that their combination contains fewer errors than each individual view. This approach is motivated by scenarios in which unconstrained text generation for error correction is too risky. We evaluated different pretrained LMs on two datasets and found significant gains in realistic scenarios with up to $15%$ WER improvement over the best OCR view. We also show the importance of domain adaptation for post-OCR correction on out-of-domain documents.

Parallel Refinements for Lexically Constrained Text Generation with BART
Xingwei He

Lexically constrained text generation aims to control the generated text by incorporating certain pre-specified keywords into the output. Previous work injects lexical constraints into the output by controlling the decoding process or refining the candidate output iteratively, which tends to generate generic or ungrammatical sentences, and has high computational complexity. To address these challenges, we proposed Constrained BART

(CBART) for lexically constrained text generation. CBART leverages the pre-trained model, BART and transfers part of the generation burden from the decoder to the encoder by decomposing this task into two sub-tasks, thereby improving the sentence quality. Concretely, we extended BART by adding a token-level classifier over the encoder, aiming at instructing the decoder where to replace and insert. Guided by the encoder, the decoder refines multiple tokens of the input in one step by inserting tokens before specific positions and re-predicting tokens at a low confidence level. To further reduce the inference latency, the decoder predicts all tokens in parallel. Experiment results on One-Billion-Word and Yelp show that CBART can generate plausible text with high quality and diversity while largely accelerating inference.

BERT-Beta: A Proactive Probabilistic Approach to Text Moderation
Fei Tan et al.

Text moderation for user generated content, which helps to promote healthy interaction among users, has been widely studied and many machine learning models have been proposed. In this work, we explore an alternative perspective by augmenting reactive reviews with proactive forecasting. Specifically, we propose a new concept text toxicity propensity to characterize the extent to which a text tends to attract toxic comments. Beta regression is then introduced to do the probabilistic modeling, which is demonstrated to function well in comprehensive experiments. We also propose an explanation method to communicate the model decision clearly. Both propensity scoring and interpretation benefit text moderation in a novel manner. Finally, the proposed scaling mechanism for the linear model offers useful insights beyond this work.

STaCK: Sentence Ordering with Temporal Commonsense Knowledge
Deepanway Ghosal et al.

Sentence order prediction is the task of finding the correct order of sentences in a randomly ordered document. Correctly ordering the sentences requires an understanding of coherence with respect to the chronological sequence of events described in the text. Document-level contextual understanding and commonsense knowledge centered around these events are often essential in uncovering this coherence and predicting the exact chronological order. In this paper, we introduce STaCK — a framework based on graph neural networks and temporal commonsense knowledge to model global information and predict the relative order of sentences. Our graph network accumulates temporal evidence using knowledge of 'past' and 'future' and formulates sentence ordering as a constrained edge classification problem. We report results on five different datasets, and empirically show that the proposed method is naturally suitable for order prediction. The implementation of this work is available at: https://github.com/declare-lab/sentence-ordering.

Preventing Author Profiling through Zero-Shot Multilingual Back-Translation
David Adelani et al.

Documents as short as a single sentence may inadvertently reveal sensitive information about their authors, including e.g. their gender or ethnicity. Style transfer is an effective way of transforming texts in order to remove any information that enables author profiling. However, for a number of current state-of-the-art approaches the improved privacy is accompanied by an undesirable drop in the down-stream utility of the transformed data. In this paper, we propose a simple, zero-shot way to effectively lower the risk of author profiling through multilingual back-translation using off-the-shelf translation models. We compare our models with five representative text style transfer models on three datasets across different domains. Results from both an automatic and a human evaluation show that our approach achieves the best overall performance while requiring no training data. We are able to lower the adversarial prediction of gender and race by up to $22%$ while retaining $95%$ of the original utility on downstream tasks.

CodeT5: Identifier-aware Unified Pre-trained Encoder-Decoder Models for Code Understanding and Generation
Yue Wang et al.

Pre-trained models for Natural Languages (NL) like BERT and GPT have been recently shown to transfer well to Programming Languages (PL) and largely benefit a broad set of code-related tasks. Despite their success, most current methods either rely on an encoder-only (or decoder-only) pre-training that is suboptimal for generation (resp. understanding) tasks or process the code snippet in the same way as NL, neglecting the special characteristics of PL such as token types. We present CodeT5, a unified pre-trained encoder-decoder Transformer model that better leverages the code semantics conveyed from the developer-assigned identifiers. Our model employs a unified framework to seamlessly support both code understanding and generation tasks and allows for multi-task learning. Besides, we propose a novel identifier-aware pre-training task that enables the model to distinguish which code tokens are identifiers and to recover them when they are masked. Furthermore, we propose to exploit the user-written code comments with a bimodal dual generation task for better NL-PL alignment. Comprehensive experiments show that CodeT5 significantly outperforms prior methods on understanding tasks such as code defect detection and clone detection, and generation tasks across various directions including PL-NL, NL-PL, and PL-PL. Further analysis reveals that our model can better capture semantic information from code. Our code and pre-trained models are released at https://github.com/salesforce/CodeT5.

Detect and Classify — Joint Span Detection and Classification for Health Outcomes
Micheal Abaho et al.

A health outcome is a measurement or an observation used to capture and assess the effect of a treatment. Automatic detection of health outcomes from text would undoubtedly speed up access to evidence necessary

in healthcare decision making. Prior work on outcome detection has modelled this task as either (a) a sequence labelling task, where the goal is to detect which text spans describe health outcomes, or (b) a classification task, where the goal is to classify a text into a predefined set of categories depending on an outcome that is mentioned somewhere in that text. However, this decoupling of span detection and classification is problematic from a modelling perspective and ignores global structural correspondences between sentence-level and word-level information present in a given text. To address this, we propose a method that uses both word-level and sentence-level information to simultaneously perform outcome span detection and outcome type classification. In addition to injecting contextual information to hidden vectors, we use label attention to appropriately weight both word and sentence level information. Experimental results on several benchmark datasets for health outcome detection show that our proposed method consistently outperforms decoupled methods, reporting competitive results.

Multi-Class Grammatical Error Detection for Correction: A Tale of Two Systems

Zheng Yuan et al.

In this paper, we show how a multi-class grammatical error detection (GED) system can be used to improve grammatical error correction (GEC) for English. Specifically, we first develop a new state-of-the-art binary detection system based on pre-trained ELECTRA, and then extend it to multi-class detection using different error type tagsets derived from the ERRANT framework. Output from this detection system is used as auxiliary input to fine-tune a novel encoder-decoder GEC model, and we subsequently re-rank the N-best GEC output to find the hypothesis that most agrees with the GED output. Results show that fine-tuning the GEC system using 4-class GED produces the best model, but re-ranking using 55-class GED leads to the best performance overall. This suggests that different multi-class GED systems benefit GEC in different ways. Ultimately, our system outperforms all other previous work that combines GED and GEC, and achieves a new single-model NMT-based state of the art on the BEA-test benchmark.

Towards Zero-shot Commonsense Reasoning with Self-supervised Refinement of Language Models

Tassilo Klein and Moin Nabi

Can we get existing language models and refine them for zero-shot commonsense reasoning? This paper presents an initial study exploring the feasibility of zero-shot commonsense reasoning for the Winograd Schema Challenge by formulating the task as self-supervised refinement of a pre-trained language model. In contrast to previous studies that rely on fine-tuning annotated datasets, we seek to boost conceptualization via loss landscape refinement. To this end, we propose a novel self-supervised learning approach that refines the language model utilizing a set of linguistic perturbations of similar concept relationships. Empirical analysis of our conceptually simple framework demonstrates the viability of zero-shot commonsense reasoning on multiple benchmarks.

To Share or not to Share: Predicting Sets of Sources for Model Transfer Learning

Lukas Lange et al.

In low-resource settings, model transfer can help to overcome a lack of labeled data for many tasks and domains. However, predicting useful transfer sources is a challenging problem, as even the most similar sources might lead to unexpected negative transfer results. Thus, ranking methods based on task and text similarity — as suggested in prior work — may not be sufficient to identify promising sources. To tackle this problem, we propose a new approach to automatically determine which and how many sources should be exploited. For this, we study the effects of model transfer on sequence labeling across various domains and tasks and show that our methods based on model similarity and support vector machines are able to predict promising sources, resulting in performance increases of up to 24 F1 points.

Self-Supervised Detection of Contextual Synonyms in a Multi-Class Setting: Phenotype Annotation Use Case

Jingqing Zhang et al.

Contextualised word embeddings is a powerful tool to detect contextual synonyms. However, most of the current state-of-the-art (SOTA) deep learning concept extraction methods remain supervised and underexploit the potential of the context. In this paper, we propose a self-supervised pre-training approach which is able to detect contextual synonyms of concepts being training on the data created by shallow matching. We apply our methodology in the sparse multi-class setting (over 15,000 concepts) to extract phenotype information from electronic health records. We further investigate data augmentation techniques to address the problem of the class sparsity. Our approach achieves a new SOTA for the unsupervised phenotype concept annotation on clinical text on F1 and Recall outperforming the previous SOTA with a gain of up to 4.5 and 4.0 absolute points, respectively. After fine-tuning with as little as 20% of the labelled data, we also outperform BioBERT and ClinicalBERT. The extrinsic evaluation on three ICU benchmarks also shows the benefit of using the phenotypes annotated by our model as features.

ClauseRec: A Clause Recommendation Framework for AI-aided Contract Authoring

Vinay Aggarwal et al.

Contracts are a common type of legal document that frequent in several day-to-day business workflows. However, there has been very limited NLP research in processing such documents, and even lesser in generating them. These contracts are made up of clauses, and the unique nature of these clauses calls for specific methods

to understand and generate such documents. In this paper, we introduce the task of clause recommendation, as a first step to aid and accelerate the authoring of contract documents. We propose a two-staged pipeline to first predict if a specific clause type is relevant to be added in a contract, and then recommend the top clauses for the given type based on the contract context. We pre-train BERT on an existing library of clauses with two additional tasks and use it for our prediction and recommendation. We experiment with classification methods and similarity-based heuristics for clause relevance prediction, and generation-based methods for clause recommendation, and evaluate the results from various methods on several clause types. We provide analyses on the results, and further outline the limitations and future directions of this line of research.

Finnish Dialect Identification: The Effect of Audio and Text
Mika Hämäläinen et al.

Finnish is a language with multiple dialects that not only differ from each other in terms of accent (pronunciation) but also in terms of morphological forms and lexical choice. We present the first approach to automatically detect the dialect of a speaker based on a dialect transcript and transcript with audio recording in a dataset consisting of 23 different dialects. Our results show that the best accuracy is received by combining both of the modalities, as text only reaches to an overall accuracy of 57%, where as text and audio reach to 85%. Our code, models and data have been released openly on Github and Zenodo.

English Machine Reading Comprehension Datasets: A Survey
Daria Dzendzik, Jennifer Foster, and Carl Vogel

This paper surveys 60 English Machine Reading Comprehension datasets, with a view to providing a convenient resource for other researchers interested in this problem. We categorize the datasets according to their question and answer form and compare them across various dimensions including size, vocabulary, data source, method of creation, human performance level, and first question word. Our analysis reveals that Wikipedia is by far the most common data source and that there is a relative lack of why, when, and where questions across datasets.

Expanding End-to-End Question Answering on Differentiable Knowledge Graphs with Intersection
Priyanka Sen, Armin Oliya, and Amir Saffari

End-to-end question answering using a differentiable knowledge graph is a promising technique that requires only weak supervision, produces interpretable results, and is fully differentiable. Previous implementations of this technique (Cohen et al, 2020) have focused on single-entity questions using a relation following operation. In this paper, we propose a model that explicitly handles multiple-entity questions by implementing a new intersection operation, which identifies the shared elements between two sets of entities. We find that introducing intersection improves performance over a baseline model on two datasets, WebQuestionsSP (69.6% to 73.3% Hits1) and ComplexWebQuestions (39.8% to 48.7% Hits1), and in particular, improves performance on questions with multiple entities by over 14% on WebQuestionsSP and by 19% on ComplexWebQuestions.

Structured Context and High-Coverage Grammar for Conversational Question Answering over Knowledge Graphs
Pierre Marion, Pawel Nowak, and Francesco Piccinno

We tackle the problem of weakly-supervised conversational Question Answering over large Knowledge Graphs using a neural semantic parsing approach. We introduce a new Logical Form (LF) grammar that can model a wide range of queries on the graph while remaining sufficiently simple to generate supervision data efficiently. Our Transformer-based model takes a JSON-like structure as input, allowing us to easily incorporate both Knowledge Graph and conversational contexts. This structured input is transformed to lists of embeddings and then fed to standard attention layers. We validate our approach, both in terms of grammar coverage and LF execution accuracy, on two publicly available datasets, CSQA and ConvQuestions, both grounded in Wikidata. On CSQA, our approach increases the coverage from 80% to 96.2%, and the LF execution accuracy from 70.6% to 75.6%, with respect to previous state-of-the-art results. On ConvQuestions, we achieve competitive results with respect to the state-of-the-art.

Improving Question Answering Model Robustness with Synthetic Adversarial Data Generation
Max Bartolo et al.

Despite recent progress, state-of-the-art question answering models remain vulnerable to a variety of adversarial attacks. While dynamic adversarial data collection, in which a human annotator tries to write examples that fool a model-in-the-loop, can improve model robustness, this process is expensive which limits the scale of the collected data. In this work, we are the first to use synthetic adversarial data generation to make question answering models more robust to human adversaries. We develop a data generation pipeline that selects source passages, identifies candidate answers, generates questions, then finally filters or re-labels them to improve quality. Using this approach, we amplify a smaller human-written adversarial dataset to a much larger set of synthetic question-answer pairs. By incorporating our synthetic data, we improve the state-of-the-art on the AdversarialQA dataset by 3.7F1 and improve model generalisation on nine of the twelve MRQA datasets. We further conduct a novel human-in-the-loop evaluation and show that our models are considerably more robust to new human-written adversarial examples: crowdworkers can fool our model only 8.8% of the time on average, compared to 17.6% for a model trained without synthetic data.

BeliefBank: Adding Memory to a Pre-Trained Language Model for a Systematic Notion of Belief
Nora Kassner et al.

Although pretrained language models (PTLMs) contain significant amounts of world knowledge, they can still produce inconsistent answers to questions when probed, even after specialized training. As a result, it can be hard to identify what the model actually "believes" about the world, making it susceptible to inconsistent behavior and simple errors. Our goal is to reduce these problems. Our approach is to embed a PTLM in a broader system that also includes an evolving, symbolic memory of beliefs – a BeliefBank – that records but then may modify the raw PTLM answers. We describe two mechanisms to improve belief consistency in the overall system. First, a reasoning component – a weighted MaxSAT solver – revises beliefs that significantly clash with others. Second, a feedback component issues future queries to the PTLM using known beliefs as context. We show that, in a controlled experimental setting, these two mechanisms result in more consistent beliefs in the overall system, improving both the accuracy and consistency of its answers over time. This is significant as it is a first step towards PTLM-based architectures with a systematic notion of belief, enabling them to construct a more coherent picture of the world, and improve over time without model retraining.

MLEC-QA: A Chinese Multi-Choice Biomedical Question Answering Dataset
Jing Li, Shangping Zhong, and Kaizhi Chen

Question Answering (QA) has been successfully applied in scenarios of human-computer interaction such as chatbots and search engines. However, for the specific biomedical domain, QA systems are still immature due to expert-annotated datasets being limited by category and scale. In this paper, we present MLEC-QA, the largest-scale Chinese multi-choice biomedical QA dataset, collected from the National Medical Licensing Examination in China. The dataset is composed of five subsets with 136,236 biomedical multi-choice questions with extra materials (images or tables) annotated by human experts, and first covers the following biomedical sub-fields: Clinic, Stomatology, Public Health, Traditional Chinese Medicine, and Traditional Chinese Medicine Combined with Western Medicine. We implement eight representative control methods and open-domain QA methods as baselines. Experimental results demonstrate that even the current best model can only achieve accuracies between 40% to 55% on five subsets, especially performing poorly on questions that require sophisticated reasoning ability. We hope the release of the MLEC-QA dataset can serve as a valuable resource for research and evaluation in open-domain QA, and also make advances for biomedical QA systems.

IndoNLG: Benchmark and Resources for Evaluating Indonesian Natural Language Generation
Samuel Cahyawijaya et al.

Natural language generation (NLG) benchmarks provide an important avenue to measure progress and develop better NLG systems. Unfortunately, the lack of publicly available NLG benchmarks for low-resource languages poses a challenging barrier for building NLG systems that work well for languages with limited amounts of data. Here we introduce IndoNLG, the first benchmark to measure natural language generation (NLG) progress in three low-resource—yet widely spoken—languages of Indonesia: Indonesian, Javanese, and Sundanese. Altogether, these languages are spoken by more than 100 million native speakers, and hence constitute an important use case of NLG systems today. Concretely, IndoNLG covers six tasks: summarization, question answering, chit-chat, and three different pairs of machine translation (MT) tasks. We collate a clean pretraining corpus of Indonesian, Sundanese, and Javanese datasets, Indo4B-Plus, which is used to pretrain our models: IndoBART and IndoGPT. We show that IndoBART and IndoGPT achieve competitive performance on all tasks—despite using only one-fifth the parameters of a larger multilingual model, mBART-large (Liu et al., 2020). This finding emphasizes the importance of pretraining on closely related, localized languages to achieve more efficient learning and faster inference at very low-resource languages like Javanese and Sundanese.

Is Multi-Hop Reasoning Really Explainable? Towards Benchmarking Reasoning Interpretability
Xin Lv et al.

Multi-hop reasoning has been widely studied in recent years to obtain more interpretable link prediction. However, we find in experiments that many paths given by these models are actually unreasonable, while little work has been done on interpretability evaluation for them. In this paper, we propose a unified framework to quantitatively evaluate the interpretability of multi-hop reasoning models so as to advance their development. In specific, we define three metrics, including path recall, local interpretability, and global interpretability for evaluation, and design an approximate strategy to calculate these metrics using the interpretability scores of rules. We manually annotate all possible rules and establish a benchmark. In experiments, we verify the effectiveness of our benchmark. Besides, we run nine representative baselines on our benchmark, and the experimental results show that the interpretability of current multi-hop reasoning models is less satisfactory and is 51.7% lower than the upper bound given by our benchmark. Moreover, the rule-based models outperform the multi-hop reasoning models in terms of performance and interpretability, which points to a direction for future research, i.e., how to better incorporate rule information into the multi-hop reasoning model. We will publish our codes and datasets upon acceptance.

Global Explainability of BERT-Based Evaluation Metrics by Disentangling along Linguistic Factors
Marvin Kaster, Wei Zhao, and Steffen Eger

Evaluation metrics are a key ingredient for progress of text generation systems. In recent years, several BERT-based evaluation metrics have been proposed (including BERTScore, MoverScore, BLEURT, etc.) which correlate much better with human assessment of text generation quality than BLEU or ROUGE, invented two decades ago. However, little is known what these metrics, which are based on black-box language model representations, actually capture (it is typically assumed they model semantic similarity). In this work, we use a simple regression based global explainability technique to disentangle metric scores along linguistic factors, including semantics, syntax, morphology, and lexical overlap. We show that the different metrics capture all aspects to some degree, but that they are all substantially sensitive to lexical overlap, just like BLEU and ROUGE. This exposes limitations of these novelly proposed metrics, which we also highlight in an adversarial test scenario.

Exploring Underexplored Limitations of Cross-Domain Text-to-SQL Generalization
Yujian Gan, Xinyun Chen, and Matthew Purver

Recently, there has been significant progress in studying neural networks for translating text descriptions into SQL queries under the zero-shot cross-domain setting. Despite achieving good performance on some public benchmarks, we observe that existing text-to-SQL models do not generalize when facing domain knowledge that does not frequently appear in the training data, which may render the worse prediction performance for unseen domains. In this work, we investigate the robustness of text-to-SQL models when the questions require rarely observed domain knowledge. In particular, we define five types of domain knowledge and introduce Spider-DK (DK is the abbreviation of domain knowledge), a human-curated dataset based on the Spider benchmark for text-to-SQL translation. NL questions in Spider-DK are selected from Spider, and we modify some samples by adding domain knowledge that reflects real-world question paraphrases. We demonstrate that the prediction accuracy dramatically drops on samples that require such domain knowledge, even if the domain knowledge appears in the training set, and the model provides the correct predictions for related training samples.

What happens if you treat ordinal ratings as interval data? Human evaluations in NLP are even more under-powered than you think
David M. Howcroft and Verena Rieser

Previous work has shown that human evaluations in NLP are notoriously under-powered. Here, we argue that there are two common factors which make this problem even worse: NLP studies usually (a) treat ordinal data as interval data and (b) operate under high variance settings while the differences they are hoping to detect are often subtle. We demonstrate through simulation that ordinal mixed effects models are better able to detect small differences between models, especially in high variance settings common in evaluations of generated texts. We release tools for researchers to conduct their own power analysis and test their assumptions. We also make recommendations for improving statistical power.

NeuTral Rewriter: A Rule-Based and Neural Approach to Automatic Rewriting into Gender Neutral Alternatives
Eva Vanmassenhove, Chris Emmery, and Dimitar Shterionov

Recent years have seen an increasing need for gender-neutral and inclusive language. Within the field of NLP, there are various mono- and bilingual use cases where gender inclusive language is appropriate, if not preferred due to ambiguity or uncertainty in terms of the gender of referents. In this work, we present a rule-based and a neural approach to gender-neutral rewriting for English along with manually curated synthetic data (WinoBias+) and natural data (OpenSubtitles and Reddit) benchmarks. A detailed manual and automatic evaluation highlights how our NeuTral Rewriter, trained on data generated by the rule-based approach, obtains word error rates (WER) below 0.18% on synthetic, in-domain and out-domain test sets.

Benchmarking Commonsense Knowledge Base Population with an Effective Evaluation Dataset
Tianqing Fang et al.

Reasoning over commonsense knowledge bases (CSKB) whose elements are in the form of free-text is an important yet hard task in NLP. While CSKB completion only fills the missing links within the domain of the CSKB, CSKB population is alternatively proposed with the goal of reasoning unseen assertions from external resources. In this task, CSKBs are grounded to a large-scale eventuality (activity, state, and event) graph to discriminate whether novel triples from the eventuality graph are plausible or not. However, existing evaluations on the population task are either not accurate (automatic evaluation with randomly sampled negative examples) or of small scale (human annotation). In this paper, we benchmark the CSKB population task with a new large-scale dataset by first aligning four popular CSKBs, and then presenting a high-quality human-annotated evaluation set to probe neural models' commonsense reasoning ability. We also propose a novel inductive commonsense reasoning model that reasons over graphs. Experimental results show that generalizing commonsense reasoning on unseen assertions is inherently a hard task. Models achieving high accuracy during training perform poorly on the evaluation set, with a large gap between human performance. We will make the data publicly available for future contributions. Codes and data are available at https://github.com/HKUST-KnowComp/CSKB-Population.

Enhancing the Context Representation in Similarity-based Word Sense Disambiguation
Ming Wang, Jianzhang Zhang, and Yinglin Wang

In previous similarity-based WSD systems, studies have allocated much effort on learning comprehensive sense embeddings using contextual representations and knowledge sources. However, the context embedding of an ambiguous word is learned using only the sentence where the word appears, neglecting its global context. In this paper, we investigate the contribution of both word-level and sense-level global context of an ambiguous word for disambiguation. Experiments have shown that the Context-Oriented Embedding (COE) can enhance a similarity-based system's performance on WSD by relatively large margins, achieving state-of-the-art on all-words WSD benchmarks in knowledge-based category.

Data Augmentation with Hierarchical SQL-to-Question Generation for Cross-domain Text-to-SQL Parsing

Kun Wu et al.

Data augmentation has attracted a lot of research attention in the deep learning era for its ability in alleviating data sparseness. The lack of labeled data for unseen evaluation databases is exactly the major challenge for cross-domain text-to-SQL parsing. Previous works either require human intervention to guarantee the quality of generated data, or fail to handle complex SQL queries. This paper presents a simple yet effective data augmentation framework. First, given a database, we automatically produce a large number of SQL queries based on an abstract syntax tree grammar. For better distribution matching, we require that at least 80% of SQL patterns in the training data are covered by generated queries. Second, we propose a hierarchical SQL-to-question generation model to obtain high-quality natural language questions, which is the major contribution of this work. Finally, we design a simple sampling strategy that can greatly improve training efficiency given large amounts of generated data. Experiments on three cross-domain datasets, i.e., WikiSQL and Spider in English, and DuSQL in Chinese, show that our proposed data augmentation framework can consistently improve performance over strong baselines, and the hierarchical generation component is the key for the improvement.

SPARQLing Database Queries from Intermediate Question Decompositions

Irina Saparina and Anton Osokin

To translate natural language questions into executable database queries, most approaches rely on a fully annotated training set. Annotating a large dataset with queries is difficult as it requires query-language expertise. We reduce this burden using grounded in databases intermediate question representations. These representations are simpler to collect and were originally crowdsourced within the Break dataset (Wolfson et al., 2020). Our pipeline consists of two parts: a neural semantic parser that converts natural language questions into the intermediate representations and a non-trainable transpiler to the SPARQL query language (a standard language for accessing knowledge graphs and semantic web). We chose SPARQL because its queries are structurally closer to our intermediate representations (compared to SQL). We observe that the execution accuracy of queries constructed by our model on the challenging Spider dataset is comparable with the state-of-the-art text-to-SQL methods trained with annotated SQL queries. Our code and data are publicly available (https://github.com/yandex-research/sparqling-queries).

Time-aware Graph Neural Network for Entity Alignment between Temporal Knowledge Graphs

Chengjin Xu, Fenglong Su, and Jens Lehmann

Entity alignment aims to identify equivalent entity pairs between different knowledge graphs (KGs). Recently, the availability of temporal KGs (TKGs) that contain time information created the need for reasoning over time in such TKGs. Existing embedding-based entity alignment approaches disregard time information that commonly exists in many large-scale KGs, leaving much room for improvement. In this paper, we focus on the task of aligning entity pairs between TKGs and propose a novel Time-aware Entity Alignment approach based on Graph Neural Networks (TEA-GNN). We embed entities, relations and timestamps of different KGs into a vector space and use GNNs to learn entity representations. To incorporate both relation and time information into the GNN structure of our model, we use a self-attention mechanism which assigns different weights to different nodes with orthogonal transformation matrices computed from embeddings of the relevant relations and timestamps in a neighborhood. Experimental results on multiple real-world TKG datasets show that our method significantly outperforms the state-of-the-art methods due to the inclusion of time information.

Cross-Domain Label-Adaptive Stance Detection

Momchil Hardalov et al.

Stance detection concerns the classification of a writer's viewpoint towards a target. There are different task variants, e.g., stance of a tweet vs. a full article, or stance with respect to a claim vs. an (implicit) topic. Moreover, task definitions vary, which includes the label inventory, the data collection, and the annotation protocol. All these aspects hinder cross-domain studies, as they require changes to standard domain adaptation approaches. In this paper, we perform an in-depth analysis of 16 stance detection datasets, and we explore the possibility for cross-domain learning from them. Moreover, we propose an end-to-end unsupervised framework for out-of-domain prediction of unseen, user-defined labels. In particular, we combine domain adaptation techniques such as mixture of experts and domain-adversarial training with label embeddings, and we demonstrate sizable performance gains over strong baselines, both (i) in-domain, i.e., for seen targets, and (ii) out-of-domain, i.e., for unseen targets. Finally, we perform an exhaustive analysis of the cross-domain results, and we highlight the important factors influencing the model performance.

Text AutoAugment: Learning Compositional Augmentation Policy for Text Classification
Shuhuai Ren et al.

Data augmentation aims to enrich training samples for alleviating the overfitting issue in low-resource or class-imbalanced situations. Traditional methods first devise task-specific operations such as Synonym Substitute, then preset the corresponding parameters such as the substitution rate artificially, which require a lot of prior knowledge and are prone to fall into the sub-optimum. Besides, the number of editing operations is limited in the previous methods, which decreases the diversity of the augmented data and thus restricts the performance gain. To overcome the above limitations, we propose a framework named Text AutoAugment (TAA) to establish a compositional and learnable paradigm for data augmentation. We regard a combination of various operations as an augmentation policy and utilize an efficient Bayesian Optimization algorithm to automatically search for the best policy, which substantially improves the generalization capability of models. Experiments on six benchmark datasets show that TAA boosts classification accuracy in low-resource and class-imbalanced regimes by an average of 8.8% and 9.7%, respectively, outperforming strong baselines.

Distilling Relation Embeddings from Pretrained Language Models
Asahi Ushio, Jose Camacho-Collados, and Steven Schockaert

Pre-trained language models have been found to capture a surprisingly rich amount of lexical knowledge, ranging from commonsense properties of everyday concepts to detailed factual knowledge about named entities. Among others, this makes it possible to distill high-quality word vectors from pre-trained language models. However, it is currently unclear to what extent it is possible to distill relation embeddings, i.e. vectors that characterize the relationship between two words. Such relation embeddings are appealing because they can, in principle, encode relational knowledge in a more fine-grained way than is possible with knowledge graphs. To obtain relation embeddings from a pre-trained language model, we encode word pairs using a (manually or automatically generated) prompt, and we fine-tune the language model such that relationally similar word pairs yield similar output vectors. We find that the resulting relation embeddings are highly competitive on analogy (unsupervised) and relation classification (supervised) benchmarks, even without any task-specific fine-tuning. Source code to reproduce our experimental results and the model checkpoints are available in the following repository: https://github.com/asahi417/relbert

Avoiding Inference Heuristics in Few-shot Prompt-based Finetuning
Prasetya Utama et al.

Recent prompt-based approaches allow pretrained language models to achieve strong performances on few-shot finetuning by reformulating downstream tasks as a language modeling problem. In this work, we demonstrate that, despite its advantages on low data regimes, finetuned prompt-based models for sentence pair classification tasks still suffer from a common pitfall of adopting inference heuristics based on lexical overlap, e.g., models incorrectly assuming a sentence pair is of the same meaning because they consist of the same set of words. Interestingly, we find that this particular inference heuristic is significantly less present in the zero-shot evaluation of the prompt-based model, indicating how finetuning can be destructive to useful knowledge learned during the pretraining. We then show that adding a regularization that preserves pretraining weights is effective in mitigating this destructive tendency of few-shot finetuning. Our evaluation on three datasets demonstrates promising improvements on the three corresponding challenge datasets used to diagnose the inference heuristics.

A Differentiable Relaxation of Graph Segmentation and Alignment for AMR Parsing
Chunchuan Lyu, Shay B. Cohen, and Ivan Titov

Abstract Meaning Representations (AMR) are a broad-coverage semantic formalism which represents sentence meaning as a directed acyclic graph. To train most AMR parsers, one needs to segment the graph into subgraphs and align each such subgraph to a word in a sentence; this is normally done at preprocessing, relying on hand-crafted rules. In contrast, we treat both alignment and segmentation as latent variables in our model and induce them as part of end-to-end training. As marginalizing over the structured latent variables is infeasible, we use the variational autoencoding framework. To ensure end-to-end differentiable optimization, we introduce a differentiable relaxation of the segmentation and alignment problems. We observe that inducing segmentation yields substantial gains over using a 'greedy' segmentation heuristic. The performance of our method also approaches that of a model that relies on the segmentation rules of Lyu and Titov (2018), which were hand-crafted to handle individual AMR constructions.

Integrating Personalized PageRank into Neural Word Sense Disambiguation
Ahmed El Sheikh, Michele Bevilacqua, and Roberto Navigli

Neural Word Sense Disambiguation (WSD) has recently been shown to benefit from the incorporation of pre-existing knowledge, such as that coming from the WordNet graph. However, state-of-the-art approaches have been successful in exploiting only the local structure of the graph, with only close neighbors of a given synset influencing the prediction. In this work, we improve a classification model by recomputing logits as a function of both the vanilla independently produced logits and the global WordNet graph. We achieve this by incorporating an online neural approximated PageRank, which enables us to refine edge weights as well. This method exploits the global graph structure while keeping space requirements linear in the number of edges. We obtain strong improvements, matching the current state of the art. Code is available at https://github.com/SapienzaNLP/neural-pagerank-wsd

Cross-lingual Sentence Embedding using Multi-Task Learning
Koustava Goswami et al.

Multilingual sentence embeddings capture rich semantic information not only for measuring similarity between texts but also for catering to a broad range of downstream cross-lingual NLP tasks. State-of-the-art multilingual sentence embedding models require large parallel corpora to learn efficiently, which confines the scope of these models. In this paper, we propose a novel sentence embedding framework based on an unsupervised loss function for generating effective multilingual sentence embeddings, eliminating the need for parallel corpora. We capture semantic similarity and relatedness between sentences using a multi-task loss function for training a dual encoder model mapping different languages onto the same vector space. We demonstrate the efficacy of an unsupervised as well as a weakly supervised variant of our framework on STS, BUCC and Tatoeba benchmark tasks. The proposed unsupervised sentence embedding framework outperforms even supervised state-of-the-art methods for certain under-resourced languages on the Tatoeba dataset and on a monolingual benchmark. Further, we show enhanced zero-shot learning capabilities for more than 30 languages, with the model being trained on only 13 languages. Our model can be extended to a wide range of languages from any language family, as it overcomes the requirement of parallel corpora for training.

NB-MLM: Efficient Domain Adaptation of Masked Language Models for Sentiment Analysis

Nikolay Arefyev, Dmitrii Kharchev, and Artem Shelmanov

While Masked Language Models (MLM) are pre-trained on massive datasets, the additional training with the MLM objective on domain or task-specific data before fine-tuning for the final task is known to improve the final performance. This is usually referred to as the domain or task adaptation step. However, unlike the initial pre-training, this step is performed for each domain or task individually and is still rather slow, requiring several GPU days compared to several GPU hours required for the final task fine-tuning. We argue that the standard MLM objective leads to inefficiency when it is used for the adaptation step because it mostly learns to predict the most frequent words, which are not necessarily related to a final task. We propose a technique for more efficient adaptation that focuses on predicting words with large weights of the Naive Bayes classifier trained for the task at hand, which are likely more relevant than the most frequent words. The proposed method provides faster adaptation and better final performance for sentiment analysis compared to the standard approach.

Revisiting Self-training for Few-shot Learning of Language Model

Yiming Chen et al.

As unlabeled data carry rich task-relevant information, they are proven useful for few-shot learning of language model. The question is how to effectively make use of such data. In this work, we revisit the self-training technique for language model fine-tuning and present a state-of-the-art prompt-based few-shot learner, SFLM. Given two views of a text sample via weak and strong augmentation techniques, SFLM generates a pseudo label on the weakly augmented version. Then, the model predicts the same pseudo label when fine-tuned with the strongly augmented version. This simple approach is shown to outperform other state-of-the-art supervised and semi-supervised counterparts on six sentence classification and six sentence-pair classification benchmarking tasks. In addition, SFLM only relies on a few in-domain unlabeled data. We conduct a comprehensive analysis to demonstrate the robustness of our proposed approach under various settings, including augmentation techniques, model scale, and few-shot knowledge transfer across tasks.

Bridging Perception, Memory, and Inference through Semantic Relations

Johanna Björklund, Adam Dahlgren Lindström, and Frank Drewes

There is a growing consensus that surface form alone does not enable models to learn meaning and gain language understanding. This warrants an interest in hybrid systems that combine the strengths of neural and symbolic methods. We favour triadic systems consisting of neural networks, knowledge bases, and inference engines. The network provides perception, that is, the interface between the system and its environment. The knowledge base provides explicit memory and thus immediate access to established facts. Finally, inference capabilities are provided by the inference engine which reflects on the perception, supported by memory, to reason and discover new facts. In this work, we probe six popular language models for semantic relations and outline a future line of research to study how the constituent subsystems can be jointly realised and integrated.

Unimodal and Crossmodal Refinement Network for Multimodal Sequence Fusion

Xiaobao Guo et al.

Effective unimodal representation and complementary crossmodal representation fusion are both important in multimodal representation learning. Prior works often modulate one modal feature to another straightforwardly and thus, underutilizing both unimodal and crossmodal representation refinements, which incurs a bottleneck of performance improvement. In this paper, Unimodal and Crossmodal Refinement Network (UCRN) is proposed to enhance both unimodal and crossmodal representations. Specifically, to improve unimodal representations, a unimodal refinement module is designed to refine modality-specific learning via iteratively updating the distribution with transformer-based attention layers. Self-quality improvement layers are followed to generate the desired weighted representations progressively. Subsequently, those unimodal representations are projected into a common latent space, regularized by a multimodal Jensen-Shannon divergence loss for better crossmodal refinement. Lastly, a crossmodal refinement module is employed to integrate all information. By hierarchical explorations on unimodal, bimodal, and trimodal interactions, UCRN is highly robust against missing modality and noisy data. Experimental results on MOSI and MOSEI datasets illustrated that the proposed UCRN outperforms recent state-of-the-art techniques and its robustness is highly preferred in real multimodal sequence fusion scenarios. Codes will be shared publicly.

YASO: A Targeted Sentiment Analysis Evaluation Dataset for Open-Domain Reviews
Matan Orbach et al.

Current TSA evaluation in a cross-domain setup is restricted to the small set of review domains available in existing datasets. Such an evaluation is limited, and may not reflect true performance on sites like Amazon or Yelp that host diverse reviews from many domains. To address this gap, we present YASO – a new TSA evaluation dataset of open-domain user reviews. YASO contains 2,215 English sentences from dozens of review domains, annotated with target terms and their sentiment. Our analysis verifies the reliability of these annotations, and explores the characteristics of the collected data. Benchmark results using five contemporary TSA systems show there is ample room for improvement on this challenging new dataset. YASO is available at https://github.com/IBM/yaso-tsa.

An Empirical Study on Leveraging Position Embeddings for Target-oriented Opinion Words Extraction
Samuel Mensah, Kai Sun, and Nikolaos Aletras

Target-oriented opinion words extraction (TOWE) (Fan et al., 2019b) is a new subtask of target-oriented sentiment analysis that aims to extract opinion words for a given aspect in text. Current state-of-the-art methods leverage position embeddings to capture the relative position of a word to the target. However, the performance of these methods depends on the ability to incorporate this information into word representations. In this paper, we explore a variety of text encoders based on pretrained word embeddings or language models that leverage part-of-speech and position embeddings, aiming to examine the actual contribution of each component in TOWE. We also adapt a graph convolutional network (GCN) to enhance word representations by incorporating syntactic information. Our experimental results demonstrate that BiLSTM-based models can effectively encode position information into word representations while using a GCN only achieves marginal gains. Interestingly, our simple methods outperform several state-of-the-art complex neural structures.

Improving Multimodal Fusion with Hierarchical Mutual Information Maximization for Multimodal Sentiment Analysis
Wei Han, Hui Chen, and Soujanya Poria

In multimodal sentiment analysis (MSA), the performance of a model highly depends on the quality of synthesized embeddings. These embeddings are generated from the upstream process called multimodal fusion, which aims to extract and combine the input unimodal raw data to produce a richer multimodal representation. Previous work either back-propagates the task loss or manipulates the geometric property of feature spaces to produce favorable fusion results, which neglects the preservation of critical task-related information that flows from input to the fusion results. In this work, we propose a framework named MultiModal InfoMax (MMIM), which hierarchically maximizes the Mutual Information (MI) in unimodal input pairs (inter-modality) and between multimodal fusion result and unimodal input in order to maintain task-related information through multimodal fusion. The framework is jointly trained with the main task (MSA) to improve the performance of the downstream MSA task. To address the intractable issue of MI bounds, we further formulate a set of computationally simple parametric and non-parametric methods to approximate their truth value. Experimental results on the two widely used datasets demonstrate the efficacy of our approach.

BERT4GCN: Using BERT Intermediate Layers to Augment GCN for Aspect-based Sentiment Classification
Zeguan Xiao et al.

Graph-based Aspect-based Sentiment Classification (ABSC) approaches have yielded state-of-the-art results, expecially when equipped with contextual word embedding from pre-training language models (PLMs). However, they ignore sequential features of the context and have not yet made the best of PLMs. In this paper, we propose a novel model, BERT4GCN, which integrates the grammatical sequential features from the PLM of BERT, and the syntactic knowledge from dependency graphs. BERT4GCN utilizes outputs from intermediate layers of BERT and positional information between words to augment GCN (Graph Convolutional Network) to better encode the dependency graphs for the downstream classification. Experimental results demonstrate that the proposed BERT4GCN outperforms all state-of-the-art baselines, justifying that augmenting GCN with the grammatical features from intermediate layers of BERT can significantly empower ABSC models.

Does Social Pressure Drive Persuasion in Online Fora?
Ayush Jain and Shashank Srivastava

Online forums such as ChangeMyView have been explored to research aspects of persuasion and argumentative quality in language. While previous research has focused on arguments between a view-holder and a persuader, we explore the premise that apart from the merits of arguments, persuasion is influenced by the ambient social community. We hypothesize that comments from the rest of the community can either affirm the original view or implicitly exert pressure to change it. We develop a structured model to capture the ambient community's sentiment towards the discussion and its effect on persuasion. Our experiments show that social features themselves are significantly predictive of persuasion (even without looking at the actual content of discussion), with performance comparable to some earlier approaches that use content features. Combining community and content features leads to overall performance of 78.5% on the persuasion prediction task. Our analyses suggest that the effect of social pressure is comparable to the difference between persuasive and non-persuasive language strategies in driving persuasion and that social pressure might be a causal factor for persuasion.

Aspect Sentiment Quad Prediction as Paraphrase Generation
Wenxuan Zhang et al.

Aspect-based sentiment analysis (ABSA) has been extensively studied in recent years, which typically involves four fundamental sentiment elements, including the aspect category, aspect term, opinion term, and sentiment polarity. Existing studies usually consider the detection of partial sentiment elements, instead of predicting the four elements in one shot. In this work, we introduce the Aspect Sentiment Quad Prediction (ASQP) task, aiming to jointly detect all sentiment elements in quads for a given opinionated sentence, which can reveal a more comprehensive and complete aspect-level sentiment structure. We further propose a novel PARAPHRASE modeling paradigm to cast the ASQP task to a paraphrase generation process. On one hand, the generation formulation allows solving ASQP in an end-to-end manner, alleviating the potential error propagation in the pipeline solution. On the other hand, the semantics of the sentiment elements can be fully exploited by learning to generate them in the natural language form. Extensive experiments on benchmark datasets show the superiority of our proposed method and the capacity of cross-task transfer with the proposed unified PARAPHRASE modeling framework.

Cross-lingual Aspect-based Sentiment Analysis with Aspect Term Code-Switching
Wenxuan Zhang et al.

Many efforts have been made in solving the Aspect-based sentiment analysis (ABSA) task. While most existing studies focus on English texts, handling ABSA in resource-poor languages remains a challenging problem. In this paper, we consider the unsupervised cross-lingual transfer for the ABSA task, where only labeled data in the source language is available and we aim at transferring its knowledge to the target language having no labeled data. To this end, we propose an alignment-free label projection method to obtain high-quality pseudo-labeled data of the target language with the help of the translation system, which could preserve more accurate task-specific knowledge in the target language. For better utilizing the source and translated data, as well as enhancing the cross-lingual alignment, we design an aspect code-switching mechanism to augment the training data with code-switched bilingual sentences. To further investigate the importance of language-specific knowledge in solving the ABSA problem, we distill the above model on the unlabeled target language data which improves the performance to the same level of the supervised method.

Towards Label-Agnostic Emotion Embeddings
Sven Buechel, Luise Modersohn, and Udo Hahn

Research in emotion analysis is scattered across different label formats (e.g., polarity types, basic emotion categories, and affective dimensions), linguistic levels (word vs. sentence vs. discourse), and, of course, (few well-resourced but much more under-resourced) natural languages and text genres (e.g., product reviews, tweets, news). The resulting heterogeneity makes data and software developed under these conflicting constraints hard to compare and challenging to integrate. To resolve this unsatisfactory state of affairs we here propose a training scheme that learns a shared latent representation of emotion independent from different label formats, natural languages, and even disparate model architectures. Experiments on a wide range of datasets indicate that this approach yields the desired interoperability without penalizing prediction quality. Code and data are archived under DOI 10.5281/zenodo.5466068.

Collaborative Learning of Bidirectional Decoders for Unsupervised Text Style Transfer
Yun Ma et al.

Unsupervised text style transfer aims to alter the underlying style of the text to a desired value while keeping its style-independent semantics, without the support of parallel training corpora. Existing methods struggle to achieve both high style conversion rate and low content loss, exhibiting the over-transfer and under-transfer problems. We attribute these problems to the conflicting driving forces of the style conversion goal and content preservation goal. In this paper, we propose a collaborative learning framework for unsupervised text style transfer using a pair of bidirectional decoders, one decoding from left to right while the other decoding from right to left. In our collaborative learning mechanism, each decoder is regularized by knowledge from its peer which has a different knowledge acquisition process. The difference is guaranteed by their opposite decoding directions and a distinguishability constraint. As a result, mutual knowledge distillation drives both decoders to a better optimum and alleviates the over-transfer and under-transfer problems. Experimental results on two benchmark datasets show that our framework achieves strong empirical results on both style compatibility and content preservation.

Exploring Non-Autoregressive Text Style Transfer
Yun Ma and Qing Li

In this paper, we explore Non-AutoRegressive (NAR) decoding for unsupervised text style transfer. We first propose a base NAR model by directly adapting the common training scheme from its AutoRegressive (AR) counterpart. Despite the faster inference speed over the AR model, this NAR model sacrifices its transfer performance due to the lack of conditional dependence between output tokens. To this end, we investigate three techniques, i.e., knowledge distillation, contrastive learning, and iterative decoding, for performance enhancement. Experimental results on two benchmark datasets suggest that, although the base NAR model is generally inferior to AR decoding, their performance gap can be clearly narrowed when empowering NAR decoding with knowledge distillation, contrastive learning, and iterative decoding.

PASTE: A Tagging-Free Decoding Framework Using Pointer Networks for Aspect Sentiment Triplet Extraction
Rajdeep Mukherjee et al.

Aspect Sentiment Triplet Extraction (ASTE) deals with extracting opinion triplets, consisting of an opinion target or aspect, its associated sentiment, and the corresponding opinion term/span explaining the rationale behind the sentiment. Existing research efforts are majorly tagging-based. Among the methods taking a sequence tagging approach, some fail to capture the strong interdependence between the three opinion factors, whereas others fall short of identifying triplets with overlapping aspect/opinion spans. A recent grid tagging approach on the other hand fails to capture the span-level semantics while predicting the sentiment between an aspect-opinion pair. Different from these, we present a tagging-free solution for the task, while addressing the limitations of the existing works. We adapt an encoder-decoder architecture with a Pointer Network-based decoding framework that generates an entire opinion triplet at each time step thereby making our solution end-to-end. Interactions between the aspects and opinions are effectively captured by the decoder by considering their entire detected spans while predicting their connecting sentiment. Extensive experiments on several benchmark datasets establish the better efficacy of our proposed approach, especially in recall, and in predicting multiple and aspect/opinion-overlapped triplets from the same review sentence. We report our results both with and without BERT and also demonstrate the utility of domain-specific BERT post-training for the task.

Adaptive Proposal Generation Network for Temporal Sentence Localization in Videos
Daizong Liu et al.

We address the problem of temporal sentence localization in videos (TSLV). Traditional methods follow a top-down framework which localizes the target segment with pre-defined segment proposals. Although they have achieved decent performance, the proposals are handcrafted and redundant. Recently, bottom-up framework attracts increasing attention due to its superior efficiency. It directly predicts the probabilities for each frame as a boundary. However, the performance of bottom-up model is inferior to the top-down counterpart as it fails to exploit the segment-level interaction. In this paper, we propose an Adaptive Proposal Generation Network (APGN) to maintain the segment-level interaction while speeding up the efficiency. Specifically, we first perform a foreground-background classification upon the video and regress on the foreground frames to adaptively generate proposals. In this way, the handcrafted proposal design is discarded and the redundant proposals are decreased. Then, a proposal consolidation module is further developed to enhance the semantics of the generated proposals. Finally, we locate the target moments with these generated proposals following the top-down framework. Extensive experiments show that our proposed APGN significantly outperforms previous state-of-the-art methods on three challenging benchmarks.

Progressively Guide to Attend: An Iterative Alignment Framework for Temporal Sentence Grounding
Daizong Liu, Xiaoye Qu, and Pan Zhou

A key solution to temporal sentence grounding (TSG) exists in how to learn effective alignment between vision and language features extracted from an untrimmed video and a sentence description. Existing methods mainly leverage vanilla soft attention to perform the alignment in a single-step process. However, such single-step attention is insufficient in practice, since complicated relations between inter- and intra-modality are usually obtained through multi-step reasoning. In this paper, we propose an Iterative Alignment Network (IA-Net) for TSG task, which iteratively interacts inter- and intra-modal features within multiple steps for more accurate grounding. Specifically, during the iterative reasoning process, we pad multi-modal features with learnable parameters to alleviate the nowhere-to-attend problem of non-matched frame-word pairs, and enhance the basic co-attention mechanism in a parallel manner. To further calibrate the misaligned attention caused by each reasoning step, we also devise a calibration module following each attention module to refine the alignment knowledge. With such iterative alignment scheme, our IA-Net can robustly capture the fine-grained relations between vision and language domains step-by-step for progressively reasoning the temporal boundaries. Extensive experiments conducted on three challenging benchmarks demonstrate that our proposed model performs better than the state-of-the-arts.

Language Models are Few-Shot Butlers
Vincent Micheli and Francois Fleuret

Pretrained language models demonstrate strong performance in most NLP tasks when fine-tuned on small task-specific datasets. Hence, these autoregressive models constitute ideal agents to operate in text-based environments where language understanding and generative capabilities are essential. Nonetheless, collecting expert demonstrations in such environments is a time-consuming endeavour. We introduce a two-stage procedure to learn from a small set of demonstrations and further improve by interacting with an environment. We show that language models fine-tuned with only 1.2% of the expert demonstrations and a simple reinforcement learning algorithm achieve a 51% absolute improvement in success rate over existing methods in the ALFWorld environment.

R3Net:Relation-embedded Representation Reconstruction Network for Change Captioning
Yunbin Tu et al.

Change captioning is to use a natural language sentence to describe the fine-grained disagreement between two similar images. Viewpoint change is the most typical distractor in this task, because it changes the scale and location of the objects and overwhelms the representation of real change. In this paper, we propose a Relation-

embedded Representation Reconstruction Network (R^3Net) to explicitly distinguish the real change from the large amount of clutter and irrelevant changes. Specifically, a relation-embedded module is first devised to explore potential changed objects in the large amount of clutter. Then, based on the semantic similarities of corresponding locations in the two images, a representation reconstruction module (RRM) is designed to learn the reconstruction representation and further model the difference representation. Besides, we introduce a syntactic skeleton predictor (SSP) to enhance the semantic interaction between change localization and caption generation. Extensive experiments show that the proposed method achieves the state-of-the-art results on two public datasets.

Looking for Confirmations: An Effective and Human-Like Visual Dialogue Strategy

Alberto Testoni and Raffaella Bernardi

Generating goal-oriented questions in Visual Dialogue tasks is a challenging and longstanding problem. State-Of-The-Art systems are shown to generate questions that, although grammatically correct, often lack an effective strategy and sound unnatural to humans. Inspired by the cognitive literature on information search and cross-situational word learning, we design Confirm-it, a model based on a beam search re-ranking algorithm that guides an effective goal-oriented strategy by asking questions that confirm the model's conjecture about the referent. We take the GuessWhat?! game as a case-study. We show that dialogues generated by Confirm-it are more natural and effective than beam search decoding without re-ranking.

A Unified Speaker Adaptation Approach for ASR

Yingzhu Zhao et al.

Transformer models have been used in automatic speech recognition (ASR) successfully and yields state-of-the-art results. However, its performance is still affected by speaker mismatch between training and test data. Further finetuning a trained model with target speaker data is the most natural approach for adaptation, but it takes a lot of compute and may cause catastrophic forgetting to the existing speakers. In this work, we propose a unified speaker adaptation approach consisting of feature adaptation and model adaptation. For feature adaptation, we employ a speaker-aware persistent memory model which generalizes better to unseen test speakers by making use of speaker i-vectors to form a persistent memory. For model adaptation, we use a novel gradual pruning method to adapt to target speakers without changing the model architecture, which to the best of our knowledge, has never been explored in ASR. Specifically, we gradually prune less contributing parameters on model encoder to a certain sparsity level, and use the pruned parameters for adaptation, while freezing the unpruned parameters to keep the original model performance. We conduct experiments on the Librispeech dataset. Our proposed approach brings relative 2.74-6.52% word error rate (WER) reduction on general speaker adaptation. On target speaker adaptation, our method outperforms the baseline with up to 20.58% relative WER reduction, and surpasses the finetuning method by up to relative 2.54%. Besides, with extremely low-resource adaptation data (e.g., 1 utterance), our method could improve the WER by relative 6.53% with only a few epochs of training.

Caption Enriched Samples for Improving Hateful Memes Detection

Efrat Blaier, Itzik Malkiel, and Lior Wolf

The recently introduced hateful meme challenge demonstrates the difficulty of determining whether a meme is hateful or not. Specifically, both unimodal language models and multimodal vision-language models cannot reach the human level of performance. Motivated by the need to model the contrast between the image content and the overlayed text, we suggest applying an off-the-shelf image captioning tool in order to capture the first. We demonstrate that the incorporation of such automatic captions during fine-tuning improves the results for various unimodal and multimodal models. Moreover, in the unimodal case, continuing the pre-training of language models on augmented and original caption pairs, is highly beneficial to the classification accuracy.

Sparsity and Sentence Structure in Encoder-Decoder Attention of Summarization Systems

Potsawee Manakul and Mark Gales

Transformer models have achieved state-of-the-art results in a wide range of NLP tasks including summarization. Training and inference using large transformer models can be computationally expensive. Previous work has focused on one important bottleneck, the quadratic self-attention mechanism in the encoder. Modified encoder architectures such as LED or LoBART use local attention patterns to address this problem for summarization. In contrast, this work focuses on the transformer's encoder-decoder attention mechanism. The cost of this attention becomes more significant in inference or training approaches that require model-generated histories. First, we examine the complexity of the encoder-decoder attention. We demonstrate empirically that there is a sparse sentence structure in document summarization that can be exploited by constraining the attention mechanism to a subset of input sentences, whilst maintaining system performance. Second, we propose a modified architecture that selects the subset of sentences to constrain the encoder-decoder attention. Experiments are carried out on abstractive summarization tasks, including CNN/DailyMail, XSum, Spotify Podcast, and arXiv.

BARThez: a Skilled Pretrained French Sequence-to-Sequence Model

Moussa Kamal Eddine, Antoine Tixier, and Michalis Vazirgiannis

Inductive transfer learning has taken the entire NLP field by storm, with models such as BERT and BART setting new state of the art on countless NLU tasks. However, most of the available models and research have been conducted for English. In this work, we introduce BARThez, the first large-scale pretrained seq2seq

model for French. Being based on BART, BARThez is particularly well-suited for generative tasks. We evaluate BARThez on five discriminative tasks from the FLUE benchmark and two generative tasks from a novel summarization dataset, OrangeSum, that we created for this research. We show BARThez to be very competitive with state-of-the-art BERT-based French language models such as CamemBERT and FlauBERT. We also continue the pretraining of a multilingual BART on BARThez' corpus, and show our resulting model, mBARThez, to significantly boost BARThez' generative performance.

ARMAN: Pre-training with Semantically Selecting and Reordering of Sentences for Persian Abstractive Summarization
Alireza Salemi et al.

Abstractive text summarization is one of the areas influenced by the emergence of pre-trained language models. Current pre-training works in abstractive summarization give more points to the summaries with more words in common with the main text and pay less attention to the semantic similarity between generated sentences and the original document. We propose ARMAN, a Transformer-based encoder-decoder model pre-trained with three novel objectives to address this issue. In ARMAN, salient sentences from a document are selected according to a modified semantic score to be masked and form a pseudo summary. To summarize more accurately and similar to human writing patterns, we applied modified sentence reordering. We evaluated our proposed models on six downstream Persian summarization tasks. Experimental results show that our proposed model achieves state-of-the-art performance on all six summarization tasks measured by ROUGE and BERTScore. Our models also outperform prior works in textual entailment, question paraphrasing, and multiple choice question answering. Finally, we established a human evaluation and show that using the semantic score significantly improves summarization results.

Models and Datasets for Cross-Lingual Summarisation
Laura Perez-Beltrachini and Mirella Lapata

We present a cross-lingual summarisation corpus with long documents in a source language associated with multi-sentence summaries in a target language. The corpus covers twelve language pairs and directions for four European languages, namely Czech, English, French and German, and the methodology for its creation can be applied to several other languages. We derive cross-lingual document-summary instances from Wikipedia by combining lead paragraphs and articles' bodies from language aligned Wikipedia titles. We analyse the proposed cross-lingual summarisation task with automatic metrics and validate it with a human study. To illustrate the utility of our dataset we report experiments with multi-lingual pre-trained models in supervised, zero- and few-shot, and out-of-domain scenarios.

Learning Opinion Summarizers by Selecting Informative Reviews
Arthur Bražinskas, Mirella Lapata, and Ivan Titov

Opinion summarization has been traditionally approached with unsupervised, weakly-supervised and few-shot learning techniques. In this work, we collect a large dataset of summaries paired with user reviews for over 31,000 products, enabling supervised training. However, the number of reviews per product is large (320 on average), making summarization — and especially training a summarizer — impractical. Moreover, the content of many reviews is not reflected in the human-written summaries, and, thus, the summarizer trained on random review subsets hallucinates. In order to deal with both of these challenges, we formulate the task as jointly learning to select informative subsets of reviews and summarizing the opinions expressed in these subsets. The choice of the review subset is treated as a latent variable, predicted by a small and simple selector. The subset is then fed into a more powerful summarizer. For joint training, we use amortized variational inference and policy gradient methods. Our experiments demonstrate the importance of selecting informative reviews resulting in improved quality of summaries and reduced hallucinations.

Enriching and Controlling Global Semantics for Text Summarization
Thong Nguyen et al.

Recently, Transformer-based models have been proven effective in the abstractive summarization task by creating fluent and informative summaries. Nevertheless, these models still suffer from the short-range dependency problem, causing them to produce summaries that miss the key points of document. In this paper, we attempt to address this issue by introducing a neural topic model empowered with normalizing flow to capture the global semantics of the document, which are then integrated into the summarization model. In addition, to avoid the overwhelming effect of global semantics on contextualized representation, we introduce a mechanism to control the amount of global semantics supplied to the text generation module. Our method outperforms state-of-the-art summarization models on five common text summarization datasets, namely CNN/DailyMail, XSum, Reddit TIFU, arXiv, and PubMed.

Revisiting Tri-training of Dependency Parsers
Joachim Wagner and Jennifer Foster

We compare two orthogonal semi-supervised learning techniques, namely tri-training and pretrained word embeddings, in the task of dependency parsing. We explore language-specific FastText and ELMo embeddings and multilingual BERT embeddings. We focus on a low resource scenario as semi-supervised learning can be expected to have the most impact here. Based on treebank size and available ELMo models, we select Hungarian, Uyghur (a zero-shot language for mBERT) and Vietnamese. Furthermore, we include English in a simulated low-resource setting. We find that pretrained word embeddings make more effective use of

unlabelled data than tri-training but that the two approaches can be successfully combined.

Bridge to Target Domain by Prototypical Contrastive Learning and Label Confusion: Re-explore Zero-Shot Learning for Slot Filling
Liwen Wang et al.

Zero-shot cross-domain slot filling alleviates the data dependence in the case of data scarcity in the target domain, which has aroused extensive research. However, as most of the existing methods do not achieve effective knowledge transfer to the target domain, they just fit the distribution of the seen slot and show poor performance on unseen slot in the target domain. To solve this, we propose a novel approach based on prototypical contrastive learning with a dynamic label confusion strategy for zero-shot slot filling. The prototypical contrastive learning aims to reconstruct the semantic constraints of labels, and we introduce the label confusion strategy to establish the label dependence between the source domains and the target domain on-the-fly. Experimental results show that our model achieves significant improvement on the unseen slots, while also set new state-of-the-arts on slot filling task.

Assessing the Reliability of Word Embedding Gender Bias Measures
Yupei Du, Qixiang Fang, and Dong Nguyen

Various measures have been proposed to quantify human-like social biases in word embeddings. However, bias scores based on these measures can suffer from measurement error. One indication of measurement quality is reliability, concerning the extent to which a measure produces consistent results. In this paper, we assess three types of reliability of word embedding gender bias measures, namely test-retest reliability, inter-rater consistency and internal consistency. Specifically, we investigate the consistency of bias scores across different choices of random seeds, scoring rules and words. Furthermore, we analyse the effects of various factors on these measures' reliability scores. Our findings inform better design of word embedding gender bias measures. Moreover, we urge researchers to be more critical about the application of such measures

SWEAT: Scoring Polarization of Topics across Different Corpora
Federico Bianchi et al.

Understanding differences of viewpoints across corpora is a fundamental task for computational social sciences. In this paper, we propose the Sliced Word Embedding Association Test (SWEAT), a novel statistical measure to compute the relative polarization of a topical wordset across two distributional representations. To this end, SWEAT uses two additional wordsets, deemed to have opposite valence, to represent two different poles. We validate our approach and illustrate a case study to show the usefulness of the introduced measure.

"So You Think You're Funny?": Rating the Humour Quotient in Standup Comedy
Anirudh Mittal et al.

Computational Humour (CH) has attracted the interest of Natural Language Processing and Computational Linguistics communities. Creating datasets for automatic measurement of humour quotient is difficult due to multiple possible interpretations of the content. In this work, we create a multi-modal humour-annotated dataset (~40 hours) using stand-up comedy clips. We devise a novel scoring mechanism to annotate the training data with a humour quotient score using the audience's laughter. The normalized duration (laughter duration divided by the clip duration) of laughter in each clip is used to compute this humour coefficient score on a five-point scale (0-4). This method of scoring is validated by comparing with manually annotated scores, wherein a quadratic weighted kappa of 0.6 is obtained. We use this dataset to train a model that provides a 'funniness' score, on a five-point scale, given the audio and its corresponding text. We compare various neural language models for the task of humour-rating and achieve an accuracy of 0.813 in terms of Quadratic Weighted Kappa (QWK). Our 'Open Mic' dataset is released for further research along with the code.

How Does Counterfactually Augmented Data Impact Models for Social Computing Constructs?
Indira Sen et al.

As NLP models are increasingly deployed in socially situated settings such as online abusive content detection, it is crucial to ensure that these models are robust. One way of improving model robustness is to generate counterfactually augmented data (CAD) for training models that can better learn to distinguish between core features and data artifacts. While models trained on this type of data have shown promising out-of-domain generalizability, it is still unclear what the sources of such improvements are. We investigate the benefits of CAD for social NLP models by focusing on three social computing constructs — sentiment, sexism, and hate speech. Assessing the performance of models trained with and without CAD across different types of datasets, we find that while models trained on CAD show lower in-domain performance, they generalize better out-of-domain. We unpack this apparent discrepancy using machine explanations and find that CAD reduces model reliance on spurious features. Leveraging a novel typology of CAD to analyze their relationship with model performance, we find that CAD which acts on the construct directly or a diverse set of CAD leads to higher performance.

Contextualize Knowledge Bases with Transformer for End-to-end Task-Oriented Dialogue Systems
Yanjie Gou et al.

Incorporating knowledge bases (KB) into end-to-end task-oriented dialogue systems is challenging, since it requires to properly represent the entity of KB, which is associated with its KB context and dialogue context. The existing works represent the entity with only perceiving a part of its KB context, which can lead to the less effective representation due to the information loss, and adversely favor KB reasoning and response generation. To tackle this issue, we explore to fully contextualize the entity representation by dynamically perceiving all the relevant entities and dialogue history. To achieve this, we propose a COntext-aware Memory Enhanced Transformer framework (COMET), which treats the KB as a sequence and leverages a novel Memory Mask to enforce the entity to only focus on its relevant entities and dialogue history, while avoiding the distraction from the irrelevant entities. Through extensive experiments, we show that our COMET framework can achieve superior performance over the state of the arts.

Efficient Dialogue Complementary Policy Learning via Deep Q-network Policy and Episodic Memory Policy
Yangyang Zhao et al.

Deep reinforcement learning has shown great potential in training dialogue policies. However, its favorable performance comes at the cost of many rounds of interaction. Most of the existing dialogue policy methods rely on a single learning system, while the human brain has two specialized learning and memory systems, supporting to find good solutions without requiring copious examples. Inspired by the human brain, this paper proposes a novel complementary policy learning (CPL) framework, which exploits the complementary advantages of the episodic memory (EM) policy and the deep Q-network (DQN) policy to achieve fast and effective dialogue policy learning. In order to coordinate between the two policies, we proposed a confidence controller to control the complementary time according to their relative efficacy at different stages. Furthermore, memory connectivity and time pruning are proposed to guarantee the flexible and adaptive generalization of the EM policy in dialog tasks. Experimental results on three dialogue datasets show that our method significantly outperforms existing methods relying on a single learning system.

We've had this conversation before: A Novel Approach to Measuring Dialog Similarity
Ofer Lavi et al.

Dialog is a core building block of human natural language interactions. It contains multi-party utterances used to convey information from one party to another in a dynamic and evolving manner. The ability to compare dialogs is beneficial in many real world use cases, such as conversation analytics for contact center calls and virtual agent design. We propose a novel adaptation of the edit distance metric to the scenario of dialog similarity. Our approach takes into account various conversation aspects such as utterance semantics, conversation flow, and the participants. We evaluate this new approach and compare it to existing document similarity measures on two publicly available datasets. The results demonstrate that our method outperforms the other approaches in capturing dialog flow, and is better aligned with the human perception of conversation similarity.

Towards Incremental Transformers: An Empirical Analysis of Transformer Models for Incremental NLU
Patrick Kahardipraja, Brielen Madureira, and David Schlangen

Incremental processing allows interactive systems to respond based on partial inputs, which is a desirable property e.g. in dialogue agents. The currently popular Transformer architecture inherently processes sequences as a whole, abstracting away the notion of time. Recent work attempts to apply Transformers incrementally via restart-incrementality by repeatedly feeding, to an unchanged model, increasingly longer input prefixes to produce partial outputs. However, this approach is computationally costly and does not scale efficiently for long sequences. In parallel, we witness efforts to make Transformers more efficient, e.g. the Linear Transformer (LT) with a recurrence mechanism. In this work, we examine the feasibility of LT for incremental NLU in English. Our results show that the recurrent LT model has better incremental performance and faster inference speed compared to the standard Transformer and LT with restart-incrementality, at the cost of part of the non-incremental (full sequence) quality. We show that the performance drop can be mitigated by training the model to wait for right context before committing to an output and that training with input prefixes is beneficial for delivering correct partial outputs.

End-to-End Learning of Flowchart Grounded Task-Oriented Dialogs
Dinesh Raghu, Shantanu Agarwal, and Sachindra Joshi

We propose a novel problem within end-to-end learning of task oriented dialogs (TOD), in which the dialog system mimics a troubleshooting agent who helps a user by diagnosing their problem (e.g., car not starting). Such dialogs are grounded in domain-specific flowcharts, which the agent is supposed to follow during the conversation. Our task exposes novel technical challenges for neural TOD, such as grounding an utterance to the flowchart without explicit annotation, referring to additional manual pages when user asks a clarification question, and ability to follow unseen flowcharts at test time. We release a dataset (FLODIAL) consisting of 2,738 dialogs grounded on 12 different troubleshooting flowcharts. We also design a neural model, FLONET, which uses a retrieval-augmented generation architecture to train the dialog agent. Our experiments find that FLONET can do zero-shot transfer to unseen flowcharts, and sets a strong baseline for future research.

Cross-lingual Intermediate Fine-tuning improves Dialogue State Tracking
Nikita Moghe, Mark Steedman, and Alexandra Birch

Recent progress in task-oriented neural dialogue systems is largely focused on a handful of languages, as annotation of training data is tedious and expensive. Machine translation has been used to make systems multilingual, but this can introduce a pipeline of errors. Another promising solution is using cross-lingual transfer learning through pretrained multilingual models. Existing methods train multilingual models with additional code-mixed task data or refine the cross-lingual representations through parallel ontologies. In this work, we enhance the transfer learning process by intermediate fine-tuning of pretrained multilingual models, where the multilingual models are fine-tuned with different but related data and/or tasks. Specifically, we use parallel and conversational movie subtitles datasets to design cross-lingual intermediate tasks suitable for downstream dialogue tasks. We use only 200K lines of parallel data for intermediate fine-tuning which is already available for 1782 language pairs. We test our approach on the cross-lingual dialogue state tracking task for the parallel MultiWoZ (English -> Chinese, Chinese -> English) and Multilingual WoZ (English -> German, English -> Italian) datasets. We achieve impressive improvements (> 20% on joint goal accuracy) on the parallel MultiWoZ dataset and the Multilingual WoZ dataset over the vanilla baseline with only 10% of the target language task data and zero-shot setup respectively.

Multilingual and Cross-Lingual Intent Detection from Spoken Data
Daniela Gerz et al.

We present a systematic study on multilingual and cross-lingual intent detection (ID) from spoken data. The study leverages a new resource put forth in this work, termed MInDS-14, a first training and evaluation resource for the ID task with spoken data. It covers 14 intents extracted from a commercial system in the e-banking domain, associated with spoken examples in 14 diverse language varieties. Our key results indicate that combining machine translation models with state-of-the-art multilingual sentence encoders (e.g., LaBSE) yield strong intent detectors in the majority of target languages covered in MInDS-14, and offer comparative analyses across different axes: e.g., translation direction, impact of speech recognition, data augmentation from a related domain. We see this work as an important step towards more inclusive development and evaluation of multilingual ID from spoken data, hopefully in a much wider spectrum of languages compared to prior work.

ConvAbuse: Data, Analysis, and Benchmarks for Nuanced Detection in Conversational AI
Amanda Cercas Curry, Gavin Abercrombie, and Verena Rieser

We present the first English corpus study on abusive language towards three conversational AI systems gathered 'in the wild': an open-domain social bot, a rule-based chatbot, and a task-based system. To account for the complexity of the task, we take a more 'nuanced' approach where our ConvAI dataset reflects fine-grained notions of abuse, as well as views from multiple expert annotators. We find that the distribution of abuse is vastly different compared to other commonly used datasets, with more sexually tinted aggression towards the virtual persona of these systems. Finally, we report results from bench-marking existing models against this data. Unsurprisingly, we find that there is substantial room for improvement with F1 scores below 90%.

Self-training Improves Pre-training for Few-shot Learning in Task-oriented Dialog Systems
Fei Mi et al.

As the labeling cost for different modules in task-oriented dialog (ToD) systems is expensive, a major challenge is to train different modules with the least amount of labeled data. Recently, large-scale pre-trained language models, have shown promising results for few-shot learning in ToD. In this paper, we devise a self-training approach to utilize the abundant unlabeled dialog data to further improve state-of-the-art pre-trained models in few-shot learning scenarios for ToD systems. Specifically, we propose a self-training approach that iteratively labels the most confident unlabeled data to train a stronger Student model. Moreover, a new text augmentation technique (GradAug) is proposed to better train the Student by replacing non-crucial tokens using a masked language model. We conduct extensive experiments and present analyses on four downstream tasks in ToD, including intent classification, dialog state tracking, dialog act prediction, and response selection. Empirical results demonstrate that the proposed self-training approach consistently improves state-of-the-art pre-trained models (BERT, ToD-BERT) when only a small number of labeled data are available.

Continual Learning in Task-Oriented Dialogue Systems
Andrea Madotto et al.

Continual learning in task-oriented dialogue systems allows the system to add new domains and functionalities overtime after deployment, without incurring the high cost of retraining the whole system each time. In this paper, we propose a first-ever continual learning benchmark for task-oriented dialogue systems with 37 domains to be learned continuously in both modularized and end-to-end learning settings. In addition, we implement and compare multiple existing continual learning baselines, and we propose a simple yet effective architectural method based on residual adapters. We also suggest that the upper bound performance of continual learning should be equivalent to multitask learning when data from all domain is available at once. Our experiments demonstrate that the proposed architectural method and a simple replay-based strategy perform better, by a large margin, compared to other continuous learning techniques, and only slightly worse than the multitask learning upper bound while being 20X faster in learning new domains. We also report several trade-offs in terms of parameter usage, memory size and training time, which are important in the design of a task-oriented dialogue system. The proposed benchmark is released to promote more research in this direction.

ConvFiT: Conversational Fine-Tuning of Pretrained Language Models
Ivan Vulić et al.

Transformer-based language models (LMs) pretrained on large text collections are proven to store a wealth of semantic knowledge. However, 1) they are not effective as sentence encoders when used off-the-shelf, and 2) thus typically lag behind conversationally pretrained (e.g., via response selection) encoders on conversational tasks such as intent detection (ID). In this work, we propose ConvFiT, a simple and efficient two-stage procedure which turns any pretrained LM into a universal conversational encoder (after Stage 1 ConvFiT-ing) and task-specialised sentence encoder (after Stage 2). We demonstrate that 1) full-blown conversational pretraining is not required, and that LMs can be quickly transformed into effective conversational encoders with much smaller amounts of unannotated data; 2) pretrained LMs can be fine-tuned into task-specialised sentence encoders, optimised for the fine-grained semantics of a particular task. Consequently, such specialised sentence encoders allow for treating ID as a simple semantic similarity task based on interpretable nearest neighbours retrieval. We validate the robustness and versatility of the ConvFiT framework with such similarity-based inference on the standard ID evaluation sets: ConvFiT-ed LMs achieve state-of-the-art ID performance across the board, with particular gains in the most challenging, few-shot setups.

Event Coreference Data (Almost) for Free: Mining Hyperlinks from Online News

Michael Bugert and Iryna Gurevych

Cross-document event coreference resolution (CDCR) is the task of identifying which event mentions refer to the same events throughout a collection of documents. Annotating CDCR data is an arduous and expensive process, explaining why existing corpora are small and lack domain coverage. To overcome this bottleneck, we automatically extract event coreference data from hyperlinks in online news: When referring to a significant real-world event, writers often add a hyperlink to another article covering this event. We demonstrate that collecting hyperlinks which point to the same article(s) produces extensive and high-quality CDCR data and create a corpus of 2M documents and 2.7M silver-standard event mentions called HyperCoref. We evaluate a state-of-the-art system on three CDCR corpora and find that models trained on small subsets of HyperCoref are highly competitive, with performance similar to models trained on gold-standard data. With our work, we free CDCR research from depending on costly human-annotated training data and open up possibilities for research beyond English CDCR, as our data extraction approach can be easily adapted to other languages.

Weakly supervised discourse segmentation for multiparty oral conversations

Lila Gravellier et al.

Discourse segmentation, the first step of discourse analysis, has been shown to improve results for text summarization, translation and other NLP tasks. While segmentation models for written text tend to perform well, they are not directly applicable to spontaneous, oral conversation, which has linguistic features foreign to written text. Segmentation is less studied for this type of language, where annotated data is scarce, and existing corpora more heterogeneous. We develop a weak supervision approach to adapt, using minimal annotation, a state of the art discourse segmenter trained on written text to French conversation transcripts. Supervision is given by a latent model bootstrapped by manually defined heuristic rules that use linguistic and acoustic information. The resulting model improves the original segmenter, especially in contexts where information on speaker turns is lacking or noisy, gaining up to 13% in F-score. Evaluation is performed on data like those used to define our heuristic rules, but also on transcripts from two other corpora.

What to Pre-Train on? Efficient Intermediate Task Selection

Clifton Poth et al.

Intermediate task fine-tuning has been shown to culminate in large transfer gains across many NLP tasks. With an abundance of candidate datasets as well as pre-trained language models, it has become infeasible to experiment with all combinations to find the best transfer setting. In this work, we provide a comprehensive comparison of different methods for efficiently identifying beneficial tasks for intermediate transfer learning. We focus on parameter and computationally efficient adapter settings, highlight different data-availability scenarios, and provide expense estimates for each method. We experiment with a diverse set of 42 intermediate and 11 target English classification, multiple choice, question answering, and sequence tagging tasks. Our results demonstrate that efficient embedding based methods, which rely solely on the respective datasets, outperform computational expensive few-shot fine-tuning approaches. Our best methods achieve an average Regret3 of 1% across all target tasks, demonstrating that we are able to efficiently identify the best datasets for intermediate training.

MATE: Multi-view Attention for Table Transformer Efficiency

Julian Eisenschlos et al.

This work presents a sparse-attention Transformer architecture for modeling documents that contain large tables. Tables are ubiquitous on the web, and are rich in information. However, more than 20% of relational tables on the web have 20 or more rows (Cafarella et al., 2008), and these large tables present a challenge for current Transformer models, which are typically limited to 512 tokens. Here we propose MATE, a novel Transformer architecture designed to model the structure of web tables. MATE uses sparse attention in a way that allows heads to efficiently attend to either rows or columns in a table. This architecture scales linearly with respect to speed and memory, and can handle documents containing more than 8000 tokens with current accelerators. MATE also has a more appropriate inductive bias for tabular data, and sets a new state-of-the-art for three table reasoning datasets. For HybridQA (Chen et al., 2020), a dataset that involves large documents containing tables, we improve the best prior result by 19 points.

Highly Parallel Autoregressive Entity Linking with Discriminative Correction
Nicola De Cao, Wilker Aziz, and Ivan Titov

Generative approaches have been recently shown to be effective for both Entity Disambiguation and Entity Linking (i.e., joint mention detection and disambiguation). However, the previously proposed autoregressive formulation for EL suffers from i) high computational cost due to a complex (deep) decoder, ii) non-parallelizable decoding that scales with the source sequence length, and iii) the need for training on a large amount of data. In this work, we propose a very efficient approach that parallelizes autoregressive linking across all potential mentions and relies on a shallow and efficient decoder. Moreover, we augment the generative objective with an extra discriminative component, i.e., a correction term which lets us directly optimize the generator's ranking. When taken together, these techniques tackle all the above issues: our model is >70 times faster and more accurate than the previous generative method, outperforming state-of-the-art approaches on the standard English dataset AIDA-CoNLL. Source code available at https://github.com/nicola-decao/efficient-autoregressive-EL

Word-Level Coreference Resolution
Vladimir Dobrovolskii

Recent coreference resolution models rely heavily on span representations to find coreference links between word spans. As the number of spans is $O(n^2)$ in the length of text and the number of potential links is $O(n^4)$, various pruning techniques are necessary to make this approach computationally feasible. We propose instead to consider coreference links between individual words rather than word spans and then reconstruct the word spans. This reduces the complexity of the coreference model to $O(n^2)$ and allows it to consider all potential mentions without pruning any of them out. We also demonstrate that, with these changes, SpanBERT for coreference resolution will be significantly outperformed by RoBERTa. While being highly efficient, our model performs competitively with recent coreference resolution systems on the OntoNotes benchmark.

Efficient Multi-Task Auxiliary Learning: Selecting Auxiliary Data by Feature Similarity
Po-Nien Kung et al.

Multi-task auxiliary learning utilizes a set of relevant auxiliary tasks to improve the performance of a primary task. A common usage is to manually select multiple auxiliary tasks for multi-task learning on all data, which raises two issues: (1) selecting beneficial auxiliary tasks for a primary task is nontrivial; (2) when the auxiliary datasets are large, training on all data becomes time-expensive and impractical. Therefore, this paper focuses on addressing these problems and proposes a time-efficient sampling method to select the data that is most relevant to the primary task. The proposed method allows us to only train on the most beneficial sub-datasets from the auxiliary tasks, achieving efficient multi-task auxiliary learning. The experiments on three benchmark datasets (RTE, MRPC, STS-B) show that our method significantly outperforms random sampling and ST-DNN. Also, by applying our method, the model can surpass fully-trained MT-DNN on RTE, MRPC, STS-B, using only 50%, 66%, and 1% of data, respectively.

Few-Shot Text Generation with Natural Language Instructions
Timo Schick and Hinrich Schütze

Providing pretrained language models with simple task descriptions in natural language enables them to solve some tasks in a fully unsupervised fashion. Moreover, when combined with regular learning from examples, this idea yields impressive few-shot results for a wide range of text classification tasks. It is also a promising direction to improve data efficiency in generative settings, but there are several challenges to using a combination of task descriptions and example-based learning for text generation. In particular, it is crucial to find task descriptions that are easy to understand for the pretrained model and to ensure that it actually makes good use of them; furthermore, effective measures against overfitting have to be implemented. In this paper, we show how these challenges can be tackled: We introduce GenPET, a method for text generation that is based on pattern-exploiting training, a recent approach for combining textual instructions with supervised learning that only works for classification tasks. On several summarization and headline generation datasets, GenPET gives consistent improvements over strong baselines in few-shot settings.

PAUSE: Positive and Annealed Unlabeled Sentence Embedding
Lele Cao et al.

Sentence embedding refers to a set of effective and versatile techniques for converting raw text into numerical vector representations that can be used in a wide range of natural language processing (NLP) applications. The majority of these techniques are either supervised or unsupervised. Compared to the unsupervised methods, the supervised ones make less assumptions about optimization objectives and usually achieve better results. However, the training requires a large amount of labeled sentence pairs, which is not available in many industrial scenarios. To that end, we propose a generic and end-to-end approach – PAUSE (Positive and Annealed Unlabeled Sentence Embedding), capable of learning high-quality sentence embeddings from a partially labeled dataset. We experimentally show that PAUSE achieves, and sometimes surpasses, state-of-the-art results using only a small fraction of labeled sentence pairs on various benchmark tasks. When applied to a real industrial use case where labeled samples are scarce, PAUSE encourages us to extend our dataset without the burden of extensive manual annotation work.

How to Train BERT with an Academic Budget
Peter Izsak, Moshe Berchansky, and Omer Levy

While large language models a la BERT are used ubiquitously in NLP, pretraining them is considered a luxury that only a few well-funded industry labs can afford. How can one train such models with a more modest budget? We present a recipe for pretraining a masked language model in 24 hours using a single low-end deep learning server. We demonstrate that through a combination of software optimizations, design choices, and hyperparameter tuning, it is possible to produce models that are competitive with BERT-base on GLUE tasks at a fraction of the original pretraining cost.

A Simple Geometric Method for Cross-Lingual Linguistic Transformations with Pre-trained Autoencoders
Maarten De Raedt et al.

Powerful sentence encoders trained for multiple languages are on the rise. These systems are capable of embedding a wide range of linguistic properties into vector representations. While explicit probing tasks can be used to verify the presence of specific linguistic properties, it is unclear whether the vector representations can be manipulated to indirectly steer such properties. For efficient learning, we investigate the use of a geometric mapping in embedding space to transform linguistic properties, without any tuning of the pre-trained sentence encoder or decoder. We validate our approach on three linguistic properties using a pre-trained multilingual autoencoder and analyze the results in both monolingual and cross-lingual settings.

Block Pruning For Faster Transformers
François Lagunas et al.

Pre-training has improved model accuracy for both classification and generation tasks at the cost of introducing much larger and slower models. Pruning methods have proven to be an effective way of reducing model size, whereas distillation methods are proven for speeding up inference. We introduce a block pruning approach targeting both small and fast models. Our approach extends structured methods by considering blocks of any size and integrates this structure into the movement pruning paradigm for fine-tuning. We find that this approach learns to prune out full components of the underlying model, such as attention heads. Experiments consider classification and generation tasks, yielding among other results a pruned model that is a 2.4x faster, 74% smaller BERT on SQuAD v1, with a 1% drop on F1, competitive both with distilled models in speed and pruned models in size.

Model Compression for Domain Adaptation through Causal Effect Estimation
Guy Rotman, Amir Feder, and Roi Reichart

Recent improvements in the predictive quality of natural language processing systems are often dependent on a substantial increase in the number of model parameters. This has led to various attempts of compressing such models, but existing methods have not considered the differences in the predictive power of various model components or in the generalizability of the compressed models. To understand the connection between model compression and out-of-distribution generalization, we define the task of compressing language representation models such that they perform best in a domain adaptation setting. We choose to address this problem from a causal perspective, attempting to estimate the average treatment effect (ATE) of a model component, such as a single layer, on the model's predictions. Our proposed ATE-guided Model Compression scheme (AMoC), generates many model candidates, differing by the model components that were removed. Then, we select the best candidate through a stepwise regression model that utilizes the ATE to predict the expected performance on the target domain. AMoC outperforms strong baselines on dozens of domain pairs across three text classification and sequence tagging tasks.

Structural Adapters in Pretrained Language Models for AMR-to-Text Generation
Leonardo F. R. Ribeiro, Yue Zhang, and Iryna Gurevych

Pretrained language models (PLM) have recently advanced graph-to-text generation, where the input graph is linearized into a sequence and fed into the PLM to obtain its representation. However, efficiently encoding the graph structure in PLMs is challenging because such models were pretrained on natural language, and modeling structured data may lead to catastrophic forgetting of distributional knowledge. In this paper, we propose StructAdapt, an adapter method to encode graph structure into PLMs. Contrary to prior work, StructAdapt effectively models interactions among the nodes based on the graph connectivity, only training graph structure-aware adapter parameters. In this way, we incorporate task-specific knowledge while maintaining the topological structure of the graph. We empirically show the benefits of explicitly encoding graph structure into PLMs using StructAdapt, outperforming the state of the art on two AMR-to-text datasets, training only 5.1% of the PLM parameters.

Smelting Gold and Silver for Improved Multilingual AMR-to-Text Generation
Leonardo F. R. Ribeiro et al.

Recent work on multilingual AMR-to-text generation has exclusively focused on data augmentation strategies that utilize silver AMR. However, this assumes a high quality of generated AMRs, potentially limiting the transferability to the target task. In this paper, we investigate different techniques for automatically generating AMR annotations, where we aim to study which source of information yields better multilingual results. Our models trained on gold AMR with silver (machine translated) sentences outperform approaches which leverage generated silver AMR. We find that combining both complementary sources of information further improves multilingual AMR-to-text generation. Our models surpass the previous state of the art for German, Italian, Spanish, and Chinese by a large margin.

Generic resources are what you need: Style transfer tasks without task-specific parallel training data
Huiyuan Lai, Antonio Toral, and Malvina Nissim

Style transfer aims to rewrite a source text in a different target style while preserving its content. We propose a novel approach to this task that leverages generic resources, and without using any task-specific parallel (source—target) data outperforms existing unsupervised approaches on the two most popular style transfer tasks: formality transfer and polarity swap. In practice, we adopt a multi-step procedure which builds on a generic pre-trained sequence-to-sequence model (BART). First, we strengthen the model's ability to rewrite by further pre-training BART on both an existing collection of generic paraphrases, as well as on synthetic pairs created using a general-purpose lexical resource. Second, through an iterative back-translation approach, we train two models, each in a transfer direction, so that they can provide each other with synthetically generated pairs, dynamically in the training process. Lastly, we let our best resulting model generate static synthetic pairs to be used in a supervised training regime. Besides methodology and state-of-the-art results, a core contribution of this work is a reflection on the nature of the two tasks we address, and how their differences are highlighted by their response to our approach.

Moral Stories: Situated Reasoning about Norms, Intents, Actions, and their Consequences
Denis Emelin et al.

In social settings, much of human behavior is governed by unspoken rules of conduct rooted in societal norms. For artificial systems to be fully integrated into social environments, adherence to such norms is a central prerequisite. To investigate whether language generation models can serve as behavioral priors for systems deployed in social settings, we evaluate their ability to generate action descriptions that achieve predefined goals under normative constraints. Moreover, we examine if models can anticipate likely consequences of actions that either observe or violate known norms, or explain why certain actions are preferable by generating relevant norm hypotheses. For this purpose, we introduce Moral Stories, a crowd-sourced dataset of structured, branching narratives for the study of grounded, goal-oriented social reasoning. Finally, we propose decoding strategies that combine multiple expert models to significantly improve the quality of generated actions, consequences, and norms compared to strong baselines.

Unsupervised Keyphrase Extraction by Jointly Modeling Local and Global Context
Xinnian Liang et al.

Embedding based methods are widely used for unsupervised keyphrase extraction (UKE) tasks. Generally, these methods simply calculate similarities between phrase embeddings and document embedding, which is insufficient to capture different context for a more effective UKE model. In this paper, we propose a novel method for UKE, where local and global contexts are jointly modeled. From a global view, we calculate the similarity between a certain phrase and the whole document in the vector space as transitional embedding based models do. In terms of the local view, we first build a graph structure based on the document where phrases are regarded as vertices and the edges are similarities between vertices. Then, we proposed a new centrality computation method to capture local salient information based on the graph structure. Finally, we further combine the modeling of global and local context for ranking. We evaluate our models on three public benchmarks (Inspec, DUC 2001, SemEval 2010) and compare with existing state-of-the-art models. The results show that our model outperforms most models while generalizing better on input documents with different domains and length. Additional ablation study shows that both the local and global information is crucial for unsupervised keyphrase extraction tasks.

Distantly Supervised Relation Extraction using Multi-Layer Revision Network and Confidence-based Multi-Instance Learning
Xiangyu Lin et al.

Distantly supervised relation extraction is widely used in the construction of knowledge bases due to its high efficiency. However, the automatically obtained instances are of low quality with numerous irrelevant words. In addition, the strong assumption of distant supervision leads to the existence of noisy sentences in the sentence bags. In this paper, we propose a novel Multi-Layer Revision Network (MLRN) which alleviates the effects of word-level noise by emphasizing inner-sentence correlations before extracting relevant information within sentences. Then, we devise a balanced and noise-resistant Confidence-based Multi-Instance Learning (CMIL) method to filter out noisy sentences as well as assign proper weights to relevant ones. Extensive experiments on two New York Times (NYT) datasets demonstrate that our approach achieves significant improvements over the baselines.

Time-dependent Entity Embedding is not All You Need: A Re-evaluation of Temporal Knowledge Graph Completion Models under a Unified Framework
Zhen Han et al.

Various temporal knowledge graph (KG) completion models have been proposed in the recent literature. The models usually contain two parts, a temporal embedding layer and a score function derived from existing static KG modeling approaches. Since the approaches differ along several dimensions, including different score functions and training strategies, the individual contributions of different temporal embedding techniques to model performance are not always clear. In this work, we systematically study six temporal embedding approaches and empirically quantify their performance across a wide range of configurations with about 3000 experiments and 13159 GPU hours. We classify the temporal embeddings into two classes: (1) timestamp embeddings and (2) time-dependent entity embeddings. Despite the common belief that the latter is more

expressive, an extensive experimental study shows that timestamp embeddings can achieve on-par or even better performance with significantly fewer parameters. Moreover, we find that when trained appropriately, the relative performance differences between various temporal embeddings often shrink and sometimes even reverse when compared to prior results. For example, TTransE leblay2018deriving, one of the first temporal KG models, can outperform more recent architectures on ICEWS datasets. To foster further research, we provide the first unified open-source framework for temporal KG completion models with full composability, where temporal embeddings, score functions, loss functions, regularizers, and the explicit modeling of reciprocal relations can be combined arbitrarily.

Uncovering Main Causalities for Long-tailed Information Extraction
Guoshun Nan et al.

Information Extraction (IE) aims to extract structural information from unstructured texts. In practice, long-tailed distributions caused by the selection bias of a dataset may lead to incorrect correlations, also known as spurious correlations, between entities and labels in the conventional likelihood models. This motivates us to propose counterfactual IE (CFIE), a novel framework that aims to uncover the main causalities behind data in the view of causal inference. Specifically, 1) we first introduce a unified structural causal model (SCM) for various IE tasks, describing the relationships among variables; 2) with our SCM, we then generate counterfactuals based on an explicit language structure to better calculate the direct causal effect during the inference stage; 3) we further propose a novel debiasing approach to yield more robust predictions. Experiments on three IE tasks across five public datasets show the effectiveness of our CFIE model in mitigating the spurious correlation issues.

Label Verbalization and Entailment for Effective Zero and Few-Shot Relation Extraction
Oscar Sainz et al.

Relation extraction systems require large amounts of labeled examples which are costly to annotate. In this work we reformulate relation extraction as an entailment task, with simple, hand-made, verbalizations of relations produced in less than 15 min per relation. The system relies on a pretrained textual entailment engine which is run as-is (no training examples, zero-shot) or further fine-tuned on labeled examples (few-shot or fully trained). In our experiments on TACRED we attain 63% F1 zero-shot, 69% with 16 examples per relation (17% points better than the best supervised system on the same conditions), and only 4 points short to the state-of-the-art (which uses 20 times more training data). We also show that the performance can be improved significantly with larger entailment models, up to 12 points in zero-shot, allowing to report the best results to date on TACRED when fully trained. The analysis shows that our few-shot systems are specially effective when discriminating between relations, and that the performance difference in low data regimes comes mainly from identifying no-relation cases.

Extend, don't rebuild: Phrasing conditional graph modification as autoregressive sequence labelling
Leon Weber et al.

Deriving and modifying graphs from natural language text has become a versatile basis technology for information extraction with applications in many subfields, such as semantic parsing or knowledge graph construction. A recent work used this technique for modifying scene graphs (He et al. 2020), by first encoding the original graph and then generating the modified one based on this encoding. In this work, we show that we can considerably increase performance on this problem by phrasing it as graph extension instead of graph generation. We propose the first model for the resulting graph extension problem based on autoregressive sequence labelling. On three scene graph modification data sets, this formulation leads to improvements in accuracy over the state-of-the-art between 13 and 24 percentage points. Furthermore, we introduce a novel data set from the biomedical domain which has much larger linguistic variability and more complex graphs than the scene graph modification data sets. For this data set, the state-of-the art fails to generalize, while our model can produce meaningful predictions.

PRIDE: Predicting Relationships in Conversations
Anna Tigunova et al.

Automatically extracting interpersonal relationships of conversation interlocutors can enrich personal knowledge bases to enhance personalized search, recommenders and chatbots. To infer speakers' relationships from dialogues we propose PRIDE, a neural multi-label classifier, based on BERT and Transformer for creating a conversation representation. PRIDE utilizes dialogue structure and augments it with external knowledge about speaker features and conversation style.Unlike prior works, we address multi-label prediction of fine-grained relationships. We release large-scale datasets, based on screenplays of movies and TV shows, with directed relationships of conversation participants. Extensive experiments on both datasets show superior performance of PRIDE compared to the state-of-the-art baselines.

Separating Retention from Extraction in the Evaluation of End-to-end Relation Extraction
Bruno Taillé et al.

State-of-the-art NLP models can adopt shallow heuristics that limit their generalization capability (McCoy et al., 2019). Such heuristics include lexical overlap with the training set in Named-Entity Recognition (Taille et al., 2020) and Event or Type heuristics in Relation Extraction (Rosenman et al., 2020). In the more realistic end-to-end RE setting, we can expect yet another heuristic: the mere retention of training relation triples. In this paper we propose two experiments confirming that retention of known facts is a key factor of performance

on standard benchmarks. Furthermore, one experiment suggests that a pipeline model able to use intermediate type representations is less prone to over-rely on retention.

Revisiting Few-shot Relation Classification: Evaluation Data and Classification Schemes
Ofer Sabo et al.

We explore Few-Shot Learning (FSL) for Relation Classification (RC). Focusing on the realistic scenario of FSL, in which a test instance might not belong to any of the target categories (none-of-the-above, aka NOTA), we first revisit the recent popular dataset structure for FSL, pointing out its unrealistic data distribution. To remedy this, we propose a novel methodology for deriving more realistic few-shot test data from available datasets for supervised RC, and apply it to the TACRED dataset. This yields a new challenging benchmark for FSL RC, on which state of the art models show poor performance. Next, we analyze classification schemes within the popular embedding-based nearest-neighbor approach for FSL, with respect to constraints they impose on the embedding space. Triggered by this analysis we propose a novel classification scheme, in which the NOTA category is represented as learned vectors, shown empirically to be an appealing option for FSL.

Monitoring geometrical properties of word embeddings for detecting the emergence of new topics.
Clément Christophe et al.

Slow emerging topic detection is a task between event detection, where we aggregate behaviors of different words on short period of time, and language evolution, where we monitor their long term evolution. In this work, we tackle the problem of early detection of slowly emerging new topics. To this end, we gather evidence of weak signals at the word level. We propose to monitor the behavior of words representation in an embedding space and use one of its geometrical properties to characterize the emergence of topics. As evaluation is typically hard for this kind of task, we present a framework for quantitative evaluation and show positive results that outperform state-of-the-art methods. Our method is evaluated on two public datasets of press and scientific articles.

IR like a SIR: Sense-enhanced Information Retrieval for Multiple Languages
Rexhina Blloshmi et al.

With the advent of contextualized embeddings, attention towards neural ranking approaches for Information Retrieval increased considerably. However, two aspects have remained largely neglected: i) queries usually consist of few keywords only, which increases ambiguity and makes their contextualization harder, and ii) performing neural ranking on non-English documents is still cumbersome due to shortage of labeled datasets. In this paper we present SIR (Sense-enhanced Information Retrieval) to mitigate both problems by leveraging word sense information. At the core of our approach lies a novel multilingual query expansion mechanism based on Word Sense Disambiguation that provides sense definitions as additional semantic information for the query. Importantly, we use senses as a bridge across languages, thus allowing our model to perform considerably better than its supervised and unsupervised alternatives across French, German, Italian and Spanish languages on several CLEF benchmarks, while being trained on English Robust04 data only. We release SIR at https://github.com/SapienzaNLP/sir.

Sociolectal Analysis of Pretrained Language Models
Sheng Zhang et al.

Using data from English cloze tests, in which subjects also self-reported their gender, age, education, and race, we examine performance differences of pretrained language models across demographic groups, defined by these (protected) attributes. We demonstrate wide performance gaps across demographic groups and show that pretrained language models systematically disfavor young non-white male speakers; i.e., not only do pretrained language models learn social biases (stereotypical associations) – pretrained language models also learn sociolectal biases, learning to speak more like some than like others. We show, however, that, with the exception of BERT models, larger pretrained language models reduce some the performance gaps between majority and minority groups.

When differential privacy meets NLP: The devil is in the detail
Ivan Habernal

Differential privacy provides a formal approach to privacy of individuals. Applications of differential privacy in various scenarios, such as protecting users' original utterances, must satisfy certain mathematical properties. Our contribution is a formal analysis of ADePT, a differentially private auto-encoder for text rewriting (Krishna et al, 2021). ADePT achieves promising results on downstream tasks while providing tight privacy guarantees. Our proof reveals that ADePT is not differentially private, thus rendering the experimental results unsubstantiated. We also quantify the impact of the error in its private mechanism, showing that the true sensitivity is higher by at least factor 6 in an optimistic case of a very small encoder's dimension and that the amount of utterances that are not privatized could easily reach 100% of the entire dataset. Our intention is neither to criticize the authors, nor the peer-reviewing process, but rather point out that if differential privacy applications in NLP rely on formal guarantees, these should be outlined in full and put under detailed scrutiny.

Achieving Model Robustness through Discrete Adversarial Training
Maor Ivgi and Jonathan Berant

Discrete adversarial attacks are symbolic perturbations to a language input that preserve the output label but lead to a prediction error. While such attacks have been extensively explored for the purpose of evaluating model robustness, their utility for improving robustness has been limited to offline augmentation only. Concretely, given a trained model, attacks are used to generate perturbed (adversarial) examples, and the model is re-trained exactly once. In this work, we address this gap and leverage discrete attacks for online augmentation, where adversarial examples are generated at every training step, adapting to the changing nature of the model. We propose (i) a new discrete attack, based on best-first search, and (ii) random sampling attacks that unlike prior work are not based on expensive search-based procedures. Surprisingly, we find that random sampling leads to impressive gains in robustness, outperforming the commonly-used offline augmentation, while leading to a speedup at training time of ~10x. Furthermore, online augmentation with search-based attacks justifies the higher training cost, significantly improving robustness on three datasets. Last, we show that our new attack substantially improves robustness compared to prior methods.

Debiasing Methods in Natural Language Understanding Make Bias More Accessible
Michael Mendelson and Yonatan Belinkov

Model robustness to bias is often determined by the generalization on carefully designed out-of-distribution datasets. Recent debiasing methods in natural language understanding (NLU) improve performance on such datasets by pressuring models into making unbiased predictions. An underlying assumption behind such methods is that this also leads to the discovery of more robust features in the model's inner representations. We propose a general probing-based framework that allows for post-hoc interpretation of biases in language models, and use an information-theoretic approach to measure the extractability of certain biases from the model's representations. We experiment with several NLU datasets and known biases, and show that, counter-intuitively, the more a language model is pushed towards a debiased regime, the more bias is actually encoded in its inner representations.

Evaluating the Robustness of Neural Language Models to Input Perturbations
Milad Moradi and Matthias Samwald

High-performance neural language models have obtained state-of-the-art results on a wide range of Natural Language Processing (NLP) tasks. However, results for common benchmark datasets often do not reflect model reliability and robustness when applied to noisy, real-world data. In this study, we design and implement various types of character-level and word-level perturbation methods to simulate realistic scenarios in which input texts may be slightly noisy or different from the data distribution on which NLP systems were trained. Conducting comprehensive experiments on different NLP tasks, we investigate the ability of high-performance language models such as BERT, XLNet, RoBERTa, and ELMo in handling different types of input perturbations. The results suggest that language models are sensitive to input perturbations and their performance can decrease even when small changes are introduced. We highlight that models need to be further improved and that current benchmarks are not reflecting model robustness well. We argue that evaluations on perturbed inputs should routinely complement widely-used benchmarks in order to yield a more realistic understanding of NLP systems' robustness.

How much pretraining data do language models need to learn syntax?
Laura Pérez-Mayos, Miguel Ballesteros, and Leo Wanner

Transformers-based pretrained language models achieve outstanding results in many well-known NLU benchmarks. However, while pretraining methods are very convenient, they are expensive in terms of time and resources. This calls for a study of the impact of pretraining data size on the knowledge of the models. We explore this impact on the syntactic capabilities of RoBERTa, using models trained on incremental sizes of raw text data. First, we use syntactic structural probes to determine whether models pretrained on more data encode a higher amount of syntactic information. Second, we perform a targeted syntactic evaluation to analyze the impact of pretraining data size on the syntactic generalization performance of the models. Third, we compare the performance of the different models on three downstream applications: part-of-speech tagging, dependency parsing and paraphrase identification. We complement our study with an analysis of the cost-benefit trade-off of training such models. Our experiments show that while models pretrained on more data encode more syntactic knowledge and perform better on downstream applications, they do not always offer a better performance across the different syntactic phenomena and come at a higher financial and environmental cost.

Examining Cross-lingual Contextual Embeddings with Orthogonal Structural Probes
Tomasz Limisiewicz and David Mareček

State-of-the-art contextual embeddings are obtained from large language models available only for a few languages. For others, we need to learn representations using a multilingual model. There is an ongoing debate on whether multilingual embeddings can be aligned in a space shared across many languages. The novel Orthogonal Structural Probe (Limisiewicz and Mareček, 2021) allows us to answer this question for specific linguistic features and learn a projection based only on mono-lingual annotated datasets. We evaluate syntactic (UD) and lexical (WordNet) structural information encoded inmBERT's contextual representations for nine diverse languages. We observe that for languages closely related to English, no transformation is needed. The evaluated information is encoded in a shared cross-lingual embedding space. For other languages, it is beneficial to apply orthogonal transformation learned separately for each language. We successfully apply our findings to zero-shot and few-shot cross-lingual parsing.

Putting Words in BERT's Mouth: Navigating Contextualized Vector Spaces with Pseudowords
Taelin Karidi et al.

We present a method for exploring regions around individual points in a contextualized vector space (particularly, BERT space), as a way to investigate how these regions correspond to word senses. By inducing a contextualized "pseudoword" vector as a stand-in for a static embedding in the input layer, and then performing masked prediction of a word in the sentence, we are able to investigate the geometry of the BERT-space in a controlled manner around individual instances. Using our method on a set of carefully constructed sentences targeting highly ambiguous English words, we find substantial regularity in the contextualized space, with regions that correspond to distinct word senses; but between these regions there are occasionally "sense voids"—regions that do not correspond to any intelligible sense.

Sorting through the noise: Testing robustness of information processing in pre-trained language models
Lalchand Pandia and Allyson Ettinger

Pre-trained LMs have shown impressive performance on downstream NLP tasks, but we have yet to establish a clear understanding of their sophistication when it comes to processing, retaining, and applying information presented in their input. In this paper we tackle a component of this question by examining robustness of models' ability to deploy relevant context information in the face of distracting content. We present models with cloze tasks requiring use of critical context information, and introduce distracting content to test how robustly the models retain and use that critical information for prediction. We also systematically manipulate the nature of these distractors, to shed light on dynamics of models' use of contextual cues. We find that although models appear in simple contexts to make predictions based on understanding and applying relevant facts from prior context, the presence of distracting but irrelevant content has clear impact in confusing model predictions. In particular, models appear particularly susceptible to factors of semantic similarity and word position. The findings are consistent with the conclusion that LM predictions are driven in large part by superficial contextual cues, rather than by robust representations of context meaning.

The Impact of Positional Encodings on Multilingual Compression
Vinit Ravishankar and Anders Søgaard

In order to preserve word-order information in a non-autoregressive setting, transformer architectures tend to include positional knowledge, by (for instance) adding positional encodings to token embeddings. Several modifications have been proposed over the sinusoidal positional encodings used in the original transformer architecture; these include, for instance, separating position encodings and token embeddings, or directly modifying attention weights based on the distance between word pairs. We first show that surprisingly, while these modifications tend to improve monolingual language models, none of them result in better multilingual language models. We then answer why that is: sinusoidal encodings were explicitly designed to facilitate compositionality by allowing linear projections over arbitrary time steps. Higher variances in multilingual training distributions requires higher compression, in which case, compositionality becomes indispensable. Learned absolute positional encodings (e.g., in mBERT) tend to approximate sinusoidal embeddings in multilingual settings, but more complex positional encoding architectures lack the inductive bias to effectively learn cross-lingual alignment. In other words, while sinusoidal positional encodings were designed for monolingual applications, they are particularly useful in multilingual language models.

Are Transformers a Modern Version of ELIZA? Observations on French Object Verb Agreement
Bingzhi Li, Guillaume Wisniewski, and Benoit Crabbé

Many recent works have demonstrated that unsupervised sentence representations of neural networks encode syntactic information by observing that neural language models are able to predict the agreement between a verb and its subject. We take a critical look at this line of research by showing that it is possible to achieve high accuracy on this agreement task with simple surface heuristics, indicating a possible flaw in our assessment of neural networks' syntactic ability. Our fine-grained analyses of results on the long-range French object-verb agreement show that contrary to LSTMs, Transformers are able to capture a non-trivial amount of grammatical structure.

Contrastive Explanations for Model Interpretability
Alon Jacovi et al.

Contrastive explanations clarify why an event occurred in contrast to another. They are inherently intuitive to humans to both produce and comprehend. We propose a method to produce contrastive explanations in the latent space, via a projection of the input representation, such that only the features that differentiate two potential decisions are captured. Our modification allows model behavior to consider only contrastive reasoning, and uncover which aspects of the input are useful for and against particular decisions. Our contrastive explanations can additionally answer for which label, and against which alternative label, is a given input feature useful. We produce contrastive explanations via both high-level abstract concept attribution and low-level input token/span attribution for two NLP classification benchmarks. Our findings demonstrate the ability of label-contrastive explanations to provide fine-grained interpretability of model decisions.

Aligning Faithful Interpretations with their Social Attribution
Alon Jacovi and Yoav Goldberg

We find that the requirement of model interpretations to be faithful is vague and incomplete. With interpretation by textual highlights as a case-study, we present several failure cases. Borrowing concepts from social science, we identify that the problem is a misalignment between the causal chain of decisions (causal attribution) and the attribution of human behavior to the interpretation (social attribution). We re-formulate faithfulness as an accurate attribution of causality to the model, and introduce the concept of aligned faithfulness: faithful causal chains that are aligned with their expected social behavior. The two steps of causal attribution and social attribution together complete the process of explaining behavior. With this formalization, we characterize various failures of misaligned faithful highlight interpretations, and propose an alternative causal chain to remedy the issues. Finally, we implement highlight explanations of the proposed causal format using contrastive explanations.

Revisiting the Uniform Information Density Hypothesis
Clara Meister et al.

The uniform information density (UID) hypothesis posits a preference among language users for utterances structured such that information is distributed uniformly across a signal. While its implications on language production have been well explored, the hypothesis potentially makes predictions about language comprehension and linguistic acceptability as well. Further, it is unclear how uniformity in a linguistic signal—or lack thereof—should be measured, and over which linguistic unit, e.g., the sentence or language level, this uniformity should hold. Here we investigate these facets of the UID hypothesis using reading time and acceptability data. While our reading time results are generally consistent with previous work, they are also consistent with a weakly super-linear effect of surprisal, which would be compatible with UID's predictions. For acceptability judgments, we find clearer evidence that non-uniformity in information density is predictive of lower acceptability. We then explore multiple operationalizations of UID, motivated by different interpretations of the original hypothesis, and analyze the scope over which the pressure towards uniformity is exerted. The explanatory power of a subset of the proposed operationalizations suggests that the strongest trend may be a regression towards a mean surprisal across the language, rather than the phrase, sentence, or document—a finding that supports a typical interpretation of UID, namely that it is the byproduct of language users maximizing the use of a (hypothetical) communication channel.

The Effect of Efficient Messaging and Input Variability on Neural-Agent Iterated Language Learning
Yuchen Lian, Arianna Bisazza, and Tessa Verhoef

Natural languages display a trade-off among different strategies to convey syntactic structure, such as word order or inflection. This trade-off, however, has not appeared in recent simulations of iterated language learning with neural network agents (Chaabouni et al., 2019b). We re-evaluate this result in light of three factors that play an important role in comparable experiments from the Language Evolution field: (i) speaker bias towards efficient messaging, (ii) non systematic input languages, and (iii) learning bottleneck. Our simulations show that neural agents mainly strive to maintain the utterance type distribution observed during learning, instead of developing a more efficient or systematic language.

Sparse Attention with Linear Units
Biao Zhang, Ivan Titov, and Rico Sennrich

Recently, it has been argued that encoder-decoder models can be made more interpretable by replacing the softmax function in the attention with its sparse variants. In this work, we introduce a novel, simple method for achieving sparsity in attention: we replace the softmax activation with a , and show that sparsity naturally emerges from such a formulation. Training stability is achieved with layer normalization with either a specialized initialization or an additional gating function. Our model, which we call Rectified Linear Attention (ReLA), is easy to implement and more efficient than previously proposed sparse attention mechanisms. We apply ReLA to the Transformer and conduct experiments on five machine translation tasks. ReLA achieves translation performance comparable to several strong baselines, with training and decoding speed similar to that of the vanilla attention. Our analysis shows that ReLA delivers high sparsity rate and head diversity, and the induced cross attention achieves better accuracy with respect to source-target word alignment than recent sparsified softmax-based models. Intriguingly, ReLA heads also learn to attend to nothing (i.e. 'switch off') for some queries, which is not possible with sparsified softmax alternatives.

Causal Direction of Data Collection Matters: Implications of Causal and Anticausal Learning for NLP
Zhijing Jin et al.

The principle of independent causal mechanisms (ICM) states that generative processes of real world data consist of independent modules which do not influence or inform each other. While this idea has led to fruitful developments in the field of causal inference, it is not widely-known in the NLP community. In this work, we argue that the causal direction of the data collection process bears nontrivial implications that can explain a number of published NLP findings, such as differences in semi-supervised learning (SSL) and domain adaptation (DA) performance across different settings. We categorize common NLP tasks according to their causal direction and empirically assay the validity of the ICM principle for text data using minimum description length. We conduct an extensive meta-analysis of over 100 published SSL and 30 DA studies, and find that the results are consistent with our expectations based on causal insights. This work presents the first attempt to analyze the ICM principle in NLP, and provides constructive suggestions for future modeling choices.

MTAdam: Automatic Balancing of Multiple Training Loss Terms
Itzik Malkiel and Lior Wolf

When training neural models, it is common to combine multiple loss terms. The balancing of these terms requires considerable human effort and is computationally demanding. Moreover, the optimal trade-off between the loss terms can change as training progresses, e.g., for adversarial terms. In this work, we generalize the Adam optimization algorithm to handle multiple loss terms. The guiding principle is that for every layer, the gradient magnitude of the terms should be balanced. To this end, the Multi-Term Adam (MTAdam) computes the derivative of each loss term separately, infers the first and second moments per parameter and loss term, and calculates a first moment for the magnitude per layer of the gradients arising from each loss. This magnitude is used to continuously balance the gradients across all layers, in a manner that both varies from one layer to the next and dynamically changes over time. Our results show that training with the new method leads to fast recovery from suboptimal initial loss weighting and to training outcomes that match or improve conventional training with the prescribed hyperparameters of each method.

KnowMAN: Weakly Supervised Multinomial Adversarial Networks
Luisa März et al.

The absence of labeled data for training neural models is often addressed by leveraging knowledge about the specific task, resulting in heuristic but noisy labels. The knowledge is captured in labeling functions, which detect certain regularities or patterns in the training samples and annotate corresponding labels for training. This process of weakly supervised training may result in an over-reliance on the signals captured by the labeling functions and hinder models to exploit other signals or to generalize well. We propose KnowMAN, an adversarial scheme that enables to control influence of signals associated with specific labeling functions. KnowMAN forces the network to learn representations that are invariant to those signals and to pick up other signals that are more generally associated with an output label. KnowMAN strongly improves results compared to direct weakly supervised learning with a pre-trained transformer language model and a feature-based baseline.

Knowledge Base Completion Meets Transfer Learning
Vid Kocijan and Thomas Lukasiewicz

The aim of knowledge base completion is to predict unseen facts from existing facts in knowledge bases. In this work, we introduce the first approach for transfer of knowledge from one collection of facts to another without the need for entity or relation matching. The method works for both canonicalized knowledge bases and uncanonicalized or open knowledge bases, i.e., knowledge bases where more than one copy of a real-world entity or relation may exist. Such knowledge bases are a natural output of automated information extraction tools that extract structured data from unstructured text. Our main contribution is a method that can make use of a large-scale pretraining on facts, collected from unstructured text, to improve predictions on structured data from a specific domain. The introduced method is the most impactful on small datasets such as ReVerb20K, where we obtained a 6% absolute increase of mean reciprocal rank and 65% relative decrease of mean rank over the previously best method, despite not relying on large pre-trained models like BERT.

Classification of hierarchical text using geometric deep learning: the case of clinical trials corpus
Sohrab Ferdowsi et al.

We consider the hierarchical representation of documents as graphs and use geometric deep learning to classify them into different categories. While graph neural networks can efficiently handle the variable structure of hierarchical documents using the permutation invariant message passing operations, we show that we can gain extra performance improvements using our proposed selective graph pooling operation that arises from the fact that some parts of the hierarchy are invariable across different documents. We applied our model to classify clinical trial (CT) protocols into completed and terminated categories. We use bag-of-words based, as well as pre-trained transformer-based embeddings to featurize the graph nodes, achieving f1-scoresaround 0.85 on a publicly available large scale CT registry of around 360K protocols. We further demonstrate how the selective pooling can add insights into the CT termination status prediction. We make the source code and dataset splits accessible.

The Devil is in the Detail: Simple Tricks Improve Systematic Generalization of Transformers
Róbert Csordás, Kazuki Irie, and Juergen Schmidhuber

Recently, many datasets have been proposed to test the systematic generalization ability of neural networks. The companion baseline Transformers, typically trained with default hyper-parameters from standard tasks, are shown to fail dramatically. Here we demonstrate that by revisiting model configurations as basic as scaling of embeddings, early stopping, relative positional embedding, and Universal Transformer variants, we can drastically improve the performance of Transformers on systematic generalization. We report improvements on five popular datasets: SCAN, CFQ, PCFG, COGS, and Mathematics dataset. Our models improve accuracy from 50% to 85% on the PCFG productivity split, and from 35% to 81% on COGS. On SCAN, relative positional embedding largely mitigates the EOS decision problem (Newman et al., 2020), yielding 100% accuracy on the length split with a cutoff at 26. Importantly, performance differences between these models are typically invisible on the IID data split. This calls for proper generalization validation sets for developing neural networks that generalize systematically. We publicly release the code to reproduce our results.

Artificial Text Detection via Examining the Topology of Attention Maps
Laida Kushnareva et al.

The impressive capabilities of recent generative models to create texts that are challenging to distinguish from the human-written ones can be misused for generating fake news, product reviews, and even abusive content. Despite the prominent performance of existing methods for artificial text detection, they still lack interpretability and robustness towards unseen models. To this end, we propose three novel types of interpretable topological features for this task based on Topological Data Analysis (TDA) which is currently understudied in the field of NLP. We empirically show that the features derived from the BERT model outperform count- and neural-based baselines up to 10% on three common datasets, and tend to be the most robust towards unseen GPT-style generation models as opposed to existing methods. The probing analysis of the features reveals their sensitivity to the surface and syntactic properties. The results demonstrate that TDA is a promising line with respect to NLP tasks, specifically the ones that incorporate surface and structural information.

SPECTRA: Sparse Structured Text Rationalization
Nuno M. Guerreiro and André F. T. Martins

Selective rationalization aims to produce decisions along with rationales (e.g., text highlights or word alignments between two sentences). Commonly, rationales are modeled as stochastic binary masks, requiring sampling-based gradient estimators, which complicates training and requires careful hyperparameter tuning. Sparse attention mechanisms are a deterministic alternative, but they lack a way to regularize the rationale extraction (e.g., to control the sparsity of a text highlight or the number of alignments). In this paper, we present a unified framework for deterministic extraction of structured explanations via constrained inference on a factor graph, forming a differentiable layer. Our approach greatly eases training and rationale regularization, generally outperforming previous work on what comes to performance and plausibility of the extracted rationales. We further provide a comparative study of stochastic and deterministic methods for rationale extraction for classification and natural language inference tasks, jointly assessing their predictive power, quality of the explanations, and model variability.

Active Learning by Acquiring Contrastive Examples
Katerina Margatina et al.

Common acquisition functions for active learning use either uncertainty or diversity sampling, aiming to select difficult and diverse data points from the pool of unlabeled data, respectively. In this work, leveraging the best of both worlds, we propose an acquisition function that opts for selecting contrastive examples, i.e. data points that are similar in the model feature space and yet the model outputs maximally different predictive likelihoods. We compare our approach, CAL (Contrastive Active Learning), with a diverse set of acquisition functions in four natural language understanding tasks and seven datasets. Our experiments show that CAL performs consistently better or equal than the best performing baseline across all tasks, on both in-domain and out-of-domain data. We also conduct an extensive ablation study of our method and we further analyze all actively acquired datasets showing that CAL achieves a better trade-off between uncertainty and diversity compared to other strategies.

Conditional Poisson Stochastic Beams
Clara Meister et al.

Beam search is the default decoding strategy for many sequence generation tasks in NLP. The set of approximate K-best items returned by the algorithm is a useful summary of the distribution for many applications; however, the candidates typically exhibit high overlap and may give a highly biased estimate for expectations under our model. These problems can be addressed by instead using stochastic decoding strategies. In this work, we propose a new method for turning beam search into a stochastic process: Conditional Poisson stochastic beam search. Rather than taking the maximizing set at each iteration, we sample K candidates without replacement according to the conditional Poisson sampling design. We view this as a more natural alternative to Kool et al. (2019)'s stochastic beam search (SBS). Furthermore, we show how samples generated under the CPSBS design can be used to build consistent estimators and sample diverse sets from sequence models. In our experiments, we observe CPSBS produces lower variance and more efficient estimators than SBS, even showing improvements in high entropy settings.

Knowledge Graph Representation Learning using Ordinary Differential Equations
Mojtaba Nayyeri et al.

Knowledge Graph Embeddings (KGEs) have shown promising performance on link prediction tasks by mapping the entities and relations from a knowledge graph into a geometric space. The capability of KGEs in preserving graph characteristics including structural aspects and semantics, highly depends on the design of their score function, as well as the inherited abilities from the underlying geometry. Many KGEs use the Euclidean geometry which renders them incapable of preserving complex structures and consequently causes wrong inferences by the models. To address this problem, we propose a neuro differential KGE that embeds nodes of a KG on the trajectories of Ordinary Differential Equations (ODEs). To this end, we represent each relation (edge) in a KG as a vector field on several manifolds. We specifically parameterize ODEs by a neural network to represent complex manifolds and complex vector fields on the manifolds. Therefore, the underlying embedding space is capable to assume the shape of various geometric forms to encode heterogeneous subgraphs. Experiments on synthetic and benchmark datasets using state-of-the-art KGE models justify the ODE trajectories as a means to enable structure preservation and consequently avoiding wrong inferences.

Measuring and Improving Consistency in Pretrained Language Models
Yanai Elazar et al.

Consistency of a model — that is, the invariance of its behavior under meaning-preserving alternations in its input — is a highly desirable property in natural language processing. In this paper we study the question: Are Pretrained Language Models (PLMs) consistent with respect to factual knowledge? To this end, we create ParaRel, a high-quality resource of cloze-style query English paraphrases. It contains a total of 328 paraphrases for 38 relations. Using ParaRel, we show that the consistency of all PLMs we experiment with is poor – though with high variance between relations. Our analysis of the representational spaces of PLMs suggests that they have a poor structure and are currently not suitable for representing knowledge robustly. Finally, we propose a method for improving model consistency and experimentally demonstrate its effectiveness.

UNKs Everywhere: Adapting Multilingual Language Models to New Scripts
Jonas Pfeiffer et al.

Massively multilingual language models such as multilingual BERT offer state-of-the-art cross-lingual transfer performance on a range of NLP tasks. However, due to limited capacity and large differences in pretraining data sizes, there is a profound performance gap between resource-rich and resource-poor target languages. The ultimate challenge is dealing with under-resourced languages not covered at all by the models and written in scripts unseen during pretraining. In this work, we propose a series of novel data-efficient methods that enable quick and effective adaptation of pretrained multilingual models to such low-resource languages and unseen scripts. Relying on matrix factorization, our methods capitalize on the existing latent knowledge about multiple languages already available in the pretrained model's embedding matrix. Furthermore, we show that learning of the new dedicated embedding matrix in the target language can be improved by leveraging a small number of vocabulary items (i.e., the so-called lexically overlapping tokens) shared between mBERT's and target language vocabulary. Our adaptation techniques offer substantial performance gains for languages with unseen scripts. We also demonstrate that they can yield improvements for low-resource languages written in scripts covered by the pretrained model.

Zero-Shot Cross-Lingual Transfer of Neural Machine Translation with Multilingual Pretrained Encoders
Guanhua Chen et al.

Previous work mainly focuses on improving cross-lingual transfer for NLU tasks with a multilingual pretrained encoder (MPE), or improving the performance on supervised machine translation with BERT. However, it is under-explored that whether the MPE can help to facilitate the cross-lingual transferability of NMT model. In this paper, we focus on a zero-shot cross-lingual transfer task in NMT. In this task, the NMT model is trained with parallel dataset of only one language pair and an off-the-shelf MPE, then it is directly tested on zero-shot language pairs. We propose SixT, a simple yet effective model for this task. SixT leverages the MPE with a two-stage training schedule and gets further improvement with a position disentangled encoder and a capacity-enhanced decoder. Using this method, SixT significantly outperforms mBART, a pretrained multilingual encoder-decoder model explicitly designed for NMT, with an average improvement of 7.1 BLEU on zero-shot any-to-English test sets across 14 source languages. Furthermore, with much less training computation cost and training data, our model achieves better performance on 15 any-to-English test sets than CRISS and m2m-100, two strong multilingual NMT baselines.

Universal Simultaneous Machine Translation with Mixture-of-Experts Wait-k Policy
Shaolei Zhang and Yang Feng

Simultaneous machine translation (SiMT) generates translation before reading the entire source sentence and hence it has to trade off between translation quality and latency. To fulfill the requirements of different translation quality and latency in practical applications, the previous methods usually need to train multiple SiMT models for different latency levels, resulting in large computational costs. In this paper, we propose a universal SiMT model with Mixture-of-Experts Wait-k Policy to achieve the best translation quality under arbitrary latency with only one trained model. Specifically, our method employs multi-head attention to accomplish the mixture of experts where each head is treated as a wait-k expert with its own waiting words number, and given a test latency and source inputs, the weights of the experts are accordingly adjusted to produce the best translation. Experiments on three datasets show that our method outperforms all the strong baselines under different latency, including the state-of-the-art adaptive policy.

Neural Machine Translation Quality and Post-Editing Performance
Vilém Zouhar et al.

We test the natural expectation that using MT in professional translation saves human processing time. The last such study was carried out by Sanchez-Torron and Koehn (2016) with phrase-based MT, artificially reducing the translation quality. In contrast, we focus on neural MT (NMT) of high quality, which has become the state-of-the-art approach since then and also got adopted by most translation companies. Through an experimental study involving over 30 professional translators for English -> Czech translation, we examine the relationship between NMT performance and post-editing time and quality. Across all models, we found that better MT systems indeed lead to fewer changes in the sentences in this industry setting. The relation between system quality and post-editing time is however not straightforward and, contrary to the results on phrase-based MT, BLEU is definitely not a stable predictor of the time or final output quality.

Speechformer: Reducing Information Loss in Direct Speech Translation
Sara Papi et al.

Transformer-based models have gained increasing popularity achieving state-of-the-art performance in many research fields including speech translation. However, Transformer's quadratic complexity with respect to the input sequence length prevents its adoption as is with audio signals, which are typically represented by long sequences. Current solutions resort to an initial sub-optimal compression based on a fixed sampling of raw audio features. Therefore, potentially useful linguistic information is not accessible to higher-level layers in the architecture. To solve this issue, we propose Speechformer, an architecture that, thanks to reduced memory usage in the attention layers, avoids the initial lossy compression and aggregates information only at a higher level according to more informed linguistic criteria. Experiments on three language pairs (en→de/es/nl) show the efficacy of our solution, with gains of up to 0.8 BLEU on the standard MuST-C corpus and of up to 4.0 BLEU in a low resource scenario.

Is "moby dick" a Whale or a Bird? Named Entities and Terminology in Speech Translation
Marco Gaido et al.

Automatic translation systems are known to struggle with rare words. Among these, named entities (NEs) and domain-specific terms are crucial, since errors in their translation can lead to severe meaning distortions. Despite their importance, previous speech translation (ST) studies have neglected them, also due to the dearth of publicly available resources tailored to their specific evaluation. To fill this gap, we i) present the first systematic analysis of the behavior of state-of-the-art ST systems in translating NEs and terminology, and ii) release NEuRoparl-ST, a novel benchmark built from European Parliament speeches annotated with NEs and terminology. Our experiments on the three language directions covered by our benchmark (en→es/fr/it) show that ST systems correctly translate 75—80% of terms and 65—70% of NEs, with very low performance (37—40%) on person names.

XTREME-R: Towards More Challenging and Nuanced Multilingual Evaluation
Sebastian Ruder et al.

Machine learning has brought striking advances in multilingual natural language processing capabilities over the past year. For example, the latest techniques have improved the state-of-the-art performance on the XTREME multilingual benchmark by more than 13 points. While a sizeable gap to human-level performance remains, improvements have been easier to achieve in some tasks than in others. This paper analyzes the current state of cross-lingual transfer learning and summarizes some lessons learned. In order to catalyze meaningful progress, we extend XTREME to XTREME-R, which consists of an improved set of ten natural language understanding tasks, including challenging language-agnostic retrieval tasks, and covers 50 typologically diverse languages. In addition, we provide a massively multilingual diagnostic suite and fine-grained multi-dataset evaluation capabilities through an interactive public leaderboard to gain a better understanding of such models.

Multilingual Unsupervised Neural Machine Translation with Denoising Adapters
Ahmet Üstün et al.

We consider the problem of multilingual unsupervised machine translation, translating to and from languages that only have monolingual data by using auxiliary parallel language pairs. For this problem the standard procedure so far to leverage the monolingual data is _back-translation_, which is computationally costly and hard to tune. In this paper we propose instead to use _denoising adapters_, adapter layers with a denoising objective, on top of pre-trained mBART-50. In addition to the modularity and flexibility of such an approach we show that the resulting translations are on-par with back-translating as measured by BLEU, and furthermore it allows adding unseen languages incrementally.

HintedBT: Augmenting Back-Translation with Quality and Transliteration Hints
Sahana Ramnath et al.

Back-translation (BT) of target monolingual corpora is a widely used data augmentation strategy for neural machine translation (NMT), especially for low-resource language pairs. To improve effectiveness of the available BT data, we introduce HintedBT—a family of techniques which provides hints (through tags) to the encoder and decoder. First, we propose a novel method of using both high and low quality BT data by providing hints (as source tags on the encoder) to the model about the quality of each source-target pair. We don't filter out low quality data but instead show that these hints enable the model to learn effectively from noisy data. Second, we address the problem of predicting whether a source token needs to be translated or transliterated to the target language, which is common in cross-script translation tasks (i.e., where source and target do not share the written script). For such cases, we propose training the model with additional hints (as target tags on the decoder) that provide information about the operation required on the source (translation or both translation and transliteration). We conduct experiments and detailed analyses on standard WMT benchmarks for three cross-script low/medium-resource language pairs: Hindi,Gujarati,Tamil-to-English. Our methods compare favorably with five strong and well established baselines. We show that using these hints, both separately and together, significantly improves translation quality and leads to state-of-the-art performance in all three language pairs in corresponding bilingual settings.

Controlling Machine Translation for Multiple Attributes with Additive Interventions
Andrea Schioppa et al.

Fine-grained control of machine translation (MT) outputs along multiple attributes is critical for many modern MT applications and is a requirement for gaining users' trust. A standard approach for exerting control in MT is to prepend the input with a special tag to signal the desired output attribute. Despite its simplicity, attribute tagging has several drawbacks: continuous values must be binned into discrete categories, which is unnatural for certain applications; interference between multiple tags is poorly understood. We address these problems by introducing vector-valued interventions which allow for fine-grained control over multiple attributes simultaneously via a weighted linear combination of the corresponding vectors. For some attributes, our approach even allows for fine-tuning a model trained without annotations to support such interventions. In experiments with three attributes (length, politeness and monotonicity) and two language pairs (English to German and Japanese) our models achieve better control over a wider range of tasks compared to tagging, and translation quality does not degrade when no control is requested. Finally, we demonstrate how to enable control in an already trained model after a relatively cheap fine-tuning stage.

A Generative Framework for Simultaneous Machine Translation
Yishu Miao, Phil Blunsom, and Lucia Specia

We propose a generative framework for simultaneous machine translation. Conventional approaches use a fixed number of source words to translate or learn dynamic policies for the number of source words by reinforcement learning. Here we formulate simultaneous translation as a structural sequence-to-sequence learning problem. A latent variable is introduced to model read or translate actions at every time step, which is then integrated out to consider all the possible translation policies. A re-parameterised Poisson prior is used to regularise the policies which allows the model to explicitly balance translation quality and latency. The experiments demonstrate the effectiveness and robustness of the generative framework, which achieves the best BLEU scores given different average translation latencies on benchmark datasets.

Nearest Neighbour Few-Shot Learning for Cross-lingual Classification
M Saiful Bari, Batool Haider, and Saab Mansour

Even though large pre-trained multilingual models (e.g. mBERT, XLM-R) have led to significant performance gains on a wide range of cross-lingual NLP tasks, success on many downstream tasks still relies on the availability of sufficient annotated data. Traditional fine-tuning of pre-trained models using only a few target samples can cause over-fitting. This can be quite limiting as most languages in the world are under-resourced. In this work, we investigate cross-lingual adaptation using a simple nearest-neighbor few-shot (<15 samples) inference technique for classification tasks. We experiment using a total of 16 distinct languages across two NLP tasks- XNLI and PAWS-X. Our approach consistently improves traditional fine-tuning using only a handful of labeled samples in target locales. We also demonstrate its generalization capability across tasks.

A Role-Selected Sharing Network for Joint Machine-Human Chatting Handoff and Service Satisfaction Analysis
Jiawei Liu et al.

Chatbot is increasingly thriving in different domains, however, because of unexpected discourse complexity and training data sparseness, its potential distrust hatches vital apprehension. Recently, Machine-Human Chatting Handoff (MHCH), predicting chatbot failure and enabling human-algorithm collaboration to enhance chatbot quality, has attracted increasing attention from industry and academia. In this study, we propose a novel model, Role-Selected Sharing Network (RSSN), which integrates both dialogue satisfaction estimation and handoff prediction in one multi-task learning framework. Unlike prior efforts in dialog mining, by utilizing local user satisfaction as a bridge, global satisfaction detector and handoff predictor can effectively exchange critical information. Specifically, we decouple the relation and interaction between the two tasks by the role information after the shared encoder. Extensive experiments on two public datasets demonstrate the effectiveness of our model.

Memory and Knowledge Augmented Language Models for Inferring Salience in Long-Form Stories
David Wilmot and Frank Keller

Measuring event salience is essential in the understanding of stories. This paper takes a recent unsupervised method for salience detection derived from Barthes Cardinal Functions and theories of surprise and applies it to longer narrative forms. We improve the standard transformer language model by incorporating an external knowledgebase (derived from Retrieval Augmented Generation) and adding a memory mechanism to enhance performance on longer works. We use a novel approach to derive salience annotation using chapter-aligned summaries from the Shmoop corpus for classic literary works. Our evaluation against this data demonstrates that our salience detection model improves performance over and above a non-knowledgebase and memory augmented language model, both of which are crucial to this improvement.

Mixture-of-Partitions: Infusing Large Biomedical Knowledge Graphs into BERT
Zaiqiao Meng et al.

Infusing factual knowledge into pre-trained models is fundamental for many knowledge-intensive tasks. In this paper, we proposed Mixture-of-Partitions (MoP), an infusion approach that can handle a very large knowledge graph (KG) by partitioning it into smaller sub-graphs and infusing their specific knowledge into various BERT models using lightweight adapters. To leverage the overall factual knowledge for a target task, these sub-graph adapters are further fine-tuned along with the underlying BERT through a mixture layer. We evaluate our

MoP with three biomedical BERTs (SciBERT, BioBERT, PubmedBERT) on six downstream tasks (inc. NLI, QA, Classification), and the results show that our MoP consistently enhances the underlying BERTs in task performance, and achieves new SOTA performances on five evaluated datasets.

MultiEURLEX - A multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer
Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos

We introduce MULTI-EURLEX, a new multilingual dataset for topic classification of legal documents. The dataset comprises 65k European Union (EU) laws, officially translated in 23 languages, annotated with multiple labels from the EUROVOC taxonomy. We highlight the effect of temporal concept drift and the importance of chronological, instead of random splits. We use the dataset as a testbed for zero-shot cross-lingual transfer, where we exploit annotated training documents in one language (source) to classify documents in another language (target). We find that fine-tuning a multilingually pretrained model (XLM-ROBERTA, MT5) in a single source language leads to catastrophic forgetting of multilingual knowledge and, consequently, poor zero-shot transfer to other languages. Adaptation strategies, namely partial fine-tuning, adapters, BITFIT, LNFIT, originally proposed to accelerate fine-tuning for new end-tasks, help retain multilingual knowledge from pretraining, substantially improving zero-shot cross-lingual transfer, but their impact also depends on the pretrained model used and the size of the label set.

Filling the Gaps in Ancient Akkadian Texts: A Masked Language Modelling Approach
Koren Lazar et al.

We present models which complete missing text given transliterations of ancient Mesopotamian documents, originally written on cuneiform clay tablets (2500 BCE - 100 CE). Due to the tablets' deterioration, scholars often rely on contextual cues to manually fill in missing parts in the text in a subjective and time-consuming process. We identify that this challenge can be formulated as a masked language modelling task, used mostly as a pretraining objective for contextualized language models. Following, we develop several architectures focusing on the Akkadian language, the lingua franca of the time. We find that despite data scarcity (1M tokens) we can achieve state of the art performance on missing tokens prediction (89% hit5) using a greedy decoding scheme and pretraining on data from other languages and different time periods. Finally, we conduct human evaluations showing the applicability of our models in assisting experts to transcribe texts in extinct languages.

CoPHE: A Count-Preserving Hierarchical Evaluation Metric in Large-Scale Multi-Label Text Classification
Matúš Falis et al.

Large-Scale Multi-Label Text Classification (LMTC) includes tasks with hierarchical label spaces, such as automatic assignment of ICD-9 codes to discharge summaries. Performance of models in prior art is evaluated with standard precision, recall, and F1 measures without regard for the rich hierarchical structure. In this work we argue for hierarchical evaluation of the predictions of neural LMTC models. With the example of the ICD-9 ontology we describe a structural issue in the representation of the structured label space in prior art, and propose an alternative representation based on the depth of the ontology. We propose a set of metrics for hierarchical evaluation using the depth-based representation. We compare the evaluation scores from the proposed metrics with previously used metrics on prior art LMTC models for ICD-9 coding in MIMIC-III. We also propose further avenues of research involving the proposed ontological representation.

Minimal Supervision for Morphological Inflection
Omer Goldman and Reut Tsarfaty

Neural models for the various flavours of morphological reinflection tasks have proven to be extremely accurate given ample labeled data, yet labeled data may be slow and costly to obtain. In this work we aim to overcome this annotation bottleneck by bootstrapping labeled data from a seed as small as *five* labeled inflection tables, accompanied by a large bulk of unlabeled text. Our bootstrapping method exploits the orthographic and semantic regularities in morphological systems in a two-phased setup, where word tagging based on *analogies* is followed by word pairing based on *distances*. Our experiments with the Paradigm Cell Filling Problem over eight typologically different languages show that in languages with relatively simple morphology, orthographic regularities on their own allow inflection models to achieve respectable accuracy. Combined orthographic and semantic regularities alleviate difficulties with particularly complex morpho-phonological systems. We further show that our bootstrapping methods substantially outperform hallucination-based methods commonly used for overcoming the annotation bottleneck in morphological reinflection tasks.

You should evaluate your language model on marginal likelihood over tokenisations
Kris Cao and Laura Rimell

Neural language models typically tokenise input text into sub-word units to achieve an open vocabulary. The standard approach is to use a single canonical tokenisation at both train and test time. We suggest that this approach is unsatisfactory and may bottleneck our evaluation of language model performance. Using only the one-best tokenisation ignores tokeniser uncertainty over alternative tokenisations, which may hurt model out-of-domain performance. In this paper, we argue that instead, language models should be evaluated on their marginal likelihood over tokenisations. We compare different estimators for the marginal likelihood based on sampling, and show that it is feasible to estimate the marginal likelihood with a manageable number

of samples. We then evaluate a pretrained language model on both the one-best-tokenisation and marginal perplexities, and show that the marginal perplexity can be significantly better than the one best, especially on out-of-domain data. We link this difference in perplexity to the tokeniser uncertainty as measured by tokeniser entropy. We discuss some implications of our results for language model training and evaluation, particularly with regard to tokenisation robustness.

Cryptonite: A Cryptic Crossword Benchmark for Extreme Ambiguity in Language
Avia Efrat et al.

Current NLP datasets targeting ambiguity can be solved by a native speaker with relative ease. We present Cryptonite, a large-scale dataset based on cryptic crosswords, which is both linguistically complex and naturally sourced. Each example in Cryptonite is a cryptic clue, a short phrase or sentence with a misleading surface reading, whose solving requires disambiguating semantic, syntactic, and phonetic wordplays, as well as world knowledge. Cryptic clues pose a challenge even for experienced solvers, though top-tier experts can solve them with almost 100% accuracy. Cryptonite is a challenging task for current models; fine-tuning T5-Large on 470k cryptic clues achieves only 7.6% accuracy, on par with the accuracy of a rule-based clue solver (8.6%).

Contrastive Domain Adaptation for Question Answering using Limited Text Corpora
Zhenrui Yue, Bernhard Kratzwald, and Stefan Feuerriegel

Question generation has recently shown impressive results in customizing question answering (QA) systems to new domains. These approaches circumvent the need for manually annotated training data from the new domain and, instead, generate synthetic question-answer pairs that are used for training. However, existing methods for question generation rely on large amounts of synthetically generated datasets and costly computational resources, which render these techniques widely inaccessible when the text corpora is of limited size. This is problematic as many niche domains rely on small text corpora, which naturally restricts the amount of synthetic data that can be generated. In this paper, we propose a novel framework for domain adaptation called contrastive domain adaptation for QA (CAQA). Specifically, CAQA combines techniques from question generation and domain-invariant learning to answer out-of-domain questions in settings with limited text corpora. Here, we train a QA system on both source data and generated data from the target domain with a contrastive adaptation loss that is incorporated in the training objective. By combining techniques from question generation and domain-invariant learning, our model achieved considerable improvements compared to state-of-the-art baselines.

Topic Transferable Table Question Answering
Saneem Chemmengath et al.

Weakly-supervised table question-answering (TableQA) models have achieved state-of-art performance by using pre-trained BERT transformer to jointly encoding a question and a table to produce structured query for the question. However, in practical settings TableQA systems are deployed over table corpora having topic and word distributions quite distinct from BERT's pretraining corpus. In this work we simulate the practical topic shift scenario by designing novel challenge benchmarks WikiSQL-TS and WikiTable-TS, consisting of train-dev-test splits in five distinct topic groups, based on the popular WikiSQL and WikiTable-Questions datasets. We empirically show that, despite pre-training on large open-domain text, performance of models degrades significantly when they are evaluated on unseen topics. In response, we propose T3QA (Topic Transferable Table Question Answering) a pragmatic adaptation framework for TableQA comprising of: (1) topic-specific vocabulary injection into BERT, (2) a novel text-to-text transformer generator (such as T5, GPT2) based natural language question generation pipeline focused on generating topic-specific training data, and (3) a logical form re-ranker. We show that T3QA provides a reasonably good baseline for our topic shift benchmarks. We believe our topic split benchmarks will lead to robust TableQA solutions that are better suited for practical deployment

Synthetic Data Augmentation for Zero-Shot Cross-Lingual Question Answering
Arij riabi et al.

Coupled with the availability of large scale datasets, deep learning architectures have enabled rapid progress on the Question Answering task. However, most of those datasets are in English, and the performances of state-of-the-art multilingual models are significantly lower when evaluated on non-English data. Due to high data collection costs, it is not realistic to obtain annotated data for each language one desires to support. We propose a method to improve the Cross-lingual Question Answering performance without requiring additional annotated data, leveraging Question Generation models to produce synthetic samples in a cross-lingual fashion. We show that the proposed method allows to significantly outperform the baselines trained on English data only. We report a new state-of-the-art on four datasets: MLQA, XQuAD, SQuAD-it and PIAF (fr).

End-to-End Entity Resolution and Question Answering Using Differentiable Knowledge Graphs
Amir Saffari et al.

Recently, end-to-end (E2E) trained models for question answering over knowledge graphs (KGQA) have delivered promising results using only a weakly supervised dataset. However, these models are trained and evaluated in a setting where hand-annotated question entities are supplied to the model, leaving the important and non-trivial task of entity resolution (ER) outside the scope of E2E learning. In this work, we extend the boundaries of E2E learning for KGQA to include the training of an ER component. Our model only needs

the question text and the answer entities to train, and delivers a stand-alone QA model that does not require an additional ER component to be supplied during runtime. Our approach is fully differentiable, thanks to its reliance on a recent method for building differentiable KGs (Cohen et al., 2020). We evaluate our E2E trained model on two public datasets and show that it comes close to baseline models that use hand-annotated entities.

Numerical reasoning in machine reading comprehension tasks: are we there yet?
Hadeel Al-Negheimish, Pranava Madhyastha, and Alessandra Russo

Numerical reasoning based machine reading comprehension is a task that involves reading comprehension along with using arithmetic operations such as addition, subtraction, sorting and counting. The DROP benchmark (Dua et al., 2019) is a recent dataset that has inspired the design of NLP models aimed at solving this task. The current standings of these models in the DROP leaderboard, over standard metrics, suggests that the models have achieved near-human performance. However, does this mean that these models have learned to reason? In this paper, we present a controlled study on some of the top-performing model architectures for the task of numerical reasoning. Our observations suggest that the standard metrics are incapable of measuring progress towards such tasks.

DWUG: A large Resource of Diachronic Word Usage Graphs in Four Languages
Dominik Schlechtweg et al.

Word meaning is notoriously difficult to capture, both synchronically and diachronically. In this paper, we describe the creation of the largest resource of graded contextualized, diachronic word meaning annotation in four different languages, based on 100,000 human semantic proximity judgments. We describe in detail the multi-round incremental annotation process, the choice for a clustering algorithm to group usages into senses, and possible — diachronic and synchronic — uses for this dataset.

I Wish I Would Have Loved This One, But I Didn't – A Multilingual Dataset for Counterfactual Detection in Product Review
James O'Neill et al.

Counterfactual statements describe events that did not or cannot take place. We consider the problem of counterfactual detection (CFD) in product reviews. For this purpose, we annotate a multilingual CFD dataset from Amazon product reviews covering counterfactual statements written in English, German, and Japanese languages. The dataset is unique as it contains counterfactuals in multiple languages, covers a new application area of e-commerce reviews, and provides high quality professional annotations. We train CFD models using different text representation methods and classifiers. We find that these models are robust against the selectional biases introduced due to cue phrase-based sentence selection. Moreover, our CFD dataset is compatible with prior datasets and can be merged to learn accurate CFD models. Applying machine translation on English counterfactual examples to create multilingual data performs poorly, demonstrating the language-specificity of this problem, which has been ignored so far.

We Need to Talk About train-dev-test Splits
Rob van der Goot

Standard train-dev-test splits used to benchmark multiple models against each other are ubiquitously used in Natural Language Processing (NLP). In this setup, the train data is used for training the model, the development set for evaluating different versions of the proposed model(s) during development, and the test set to confirm the answers to the main research question(s). However, the introduction of neural networks in NLP has led to a different use of these standard splits; the development set is now often used for model selection during the training procedure. Because of this, comparing multiple versions of the same model during development leads to overestimation on the development data. As an effect, people have started to compare an increasing amount of models on the test data, leading to faster overfitting and "expiration" of our test sets. We propose to use a tune-set when developing neural network methods, which can be used for model picking so that comparing the different versions of a new model can safely be done on the development data.

Does It Capture STEL? A Modular, Similarity-based Linguistic Style Evaluation Framework
Anna Wegmann and Dong Nguyen

Style is an integral part of natural language. However, evaluation methods for style measures are rare, often task-specific and usually do not control for content. We propose the modular, fine-grained and content-controlled similarity-based STyle EvaLuation framework (STEL) to test the performance of any model that can compare two sentences on style. We illustrate STEL with two general dimensions of style (formal/informal and simple/complex) as well as two specific characteristics of style (contrac'tion and numb3r substitution). We find that BERT-based methods outperform simple versions of commonly used style measures like 3-grams, punctuation frequency and LIWC-based approaches. We invite the addition of further tasks and task instances to STEL and hope to facilitate the improvement of style-sensitive measures.

A Large-Scale Dataset for Empathetic Response Generation
Anuradha Welivita, Yubo Xie, and Pearl Pu

Recent development in NLP shows a strong trend towards refining pre-trained models with a domain-specific dataset. This is especially the case for response generation where emotion plays an important role. However,

existing empathetic datasets remain small, delaying research efforts in this area, for example, the development of emotion-aware chatbots. One main technical challenge has been the cost of manually annotating dialogues with the right emotion labels. In this paper, we describe a large-scale silver dataset consisting of 1M dialogues annotated with 32 fine-grained emotions, eight empathetic response intents, and the Neutral category. To achieve this goal, we have developed a novel data curation pipeline starting with a small seed of manually annotated data and eventually scaling it to a satisfactory size. We compare its quality against a state-of-the-art gold dataset using both offline experiments and visual validation methods. The resultant procedure can be used to create similar datasets in the same domain as well as in other domains.

Automatic Text Evaluation through the Lens of Wasserstein Barycenters
Pierre Colombo et al.

A new metric `BaryScore` to evaluate text generation based on deep contextualized embeddings (*e.g.*, BERT, Roberta, ELMo) is introduced. This metric is motivated by a new framework relying on optimal transport tools, *i.e.*, Wasserstein distance and barycenter. By modelling the layer output of deep contextualized embeddings as a probability distribution rather than by a vector embedding; this framework provides a natural way to aggregate the different outputs through the Wasserstein space topology. In addition, it provides theoretical grounds to our metric and offers an alternative to available solutions (*e.g.*, MoverScore and BertScore). Numerical evaluation is performed on four different tasks: machine translation, summarization, data2text generation and image captioning. Our results show that `BaryScore` outperforms other BERT based metrics and exhibits more consistent behaviour in particular for text summarization.

AM2iCo: Evaluating Word Meaning in Context across Low-Resource Languages with Adversarial Examples
Qianchu Liu et al.

Capturing word meaning in context and distinguishing between correspondences and variations across languages is key to building successful multilingual and cross-lingual text representation models. However, existing multilingual evaluation datasets that evaluate lexical semantics "in-context" have various limitations. In particular, 1) their language coverage is restricted to high-resource languages and skewed in favor of only a few language families and areas, 2) a design that makes the task solvable via superficial cues, which results in artificially inflated (and sometimes super-human) performances of pretrained encoders, and 3) no support for cross-lingual evaluation. In order to address these gaps, we present AM2iCo (Adversarial and Multilingual Meaning in Context), a wide-coverage cross-lingual and multilingual evaluation set; it aims to faithfully assess the ability of state-of-the-art (SotA) representation models to understand the identity of word meaning in cross-lingual contexts for 14 language pairs. We conduct a series of experiments in a wide range of setups and demonstrate the challenging nature of AM2iCo. The results reveal that current SotA pretrained encoders substantially lag behind human performance, and the largest gaps are observed for low-resource languages and languages dissimilar to English.

Visually Grounded Reasoning across Languages and Cultures
Fangyu Liu et al.

The design of widespread vision-and-language datasets and pre-trained encoders directly adopts, or draws inspiration from, the concepts and images of ImageNet. While one can hardly overestimate how much this benchmark contributed to progress in computer vision, it is mostly derived from lexical databases and image queries in English, resulting in source material with a North American or Western European bias. Therefore, we devise a new protocol to construct an ImageNet-style hierarchy representative of more languages and cultures. In particular, we let the selection of both concepts and images be entirely driven by native speakers, rather than scraping them automatically. Specifically, we focus on a typologically diverse set of languages, namely, Indonesian, Mandarin Chinese, Swahili, Tamil, and Turkish. On top of the concepts and images obtained through this new protocol, we create a multilingual dataset for Multicultural Reasoning over Vision and Language (MaRVL) by eliciting statements from native speaker annotators about pairs of images. The task consists of discriminating whether each grounded statement is true or false. We establish a series of baselines using state-of-the-art models and find that their cross-lingual transfer performance lags dramatically behind supervised performance in English. These results invite us to reassess the robustness and accuracy of current state-of-the-art models beyond a narrow domain, but also open up new exciting challenges for the development of truly multilingual and multicultural systems.

CSDS: A Fine-Grained Chinese Dataset for Customer Service Dialogue Summarization
Haitao Lin et al.

Dialogue summarization has drawn much attention recently. Especially in the customer service domain, agents could use dialogue summaries to help boost their works by quickly knowing customer's issues and service progress. These applications require summaries to contain the perspective of a single speaker and have a clear topic flow structure, while neither are available in existing datasets. Therefore, in this paper, we introduce a novel Chinese dataset for Customer Service Dialogue Summarization (CSDS). CSDS improves the abstractive summaries in two aspects: (1) In addition to the overall summary for the whole dialogue, role-oriented summaries are also provided to acquire different speakers' viewpoints. (2) All the summaries sum up each topic separately, thus containing the topic-level structure of the dialogue. We define tasks in CSDS as generating the overall summary and different role-oriented summaries for a given dialogue. Next, we compare various summarization methods on CSDS, and experiment results show that existing methods are prone to generate redundant and incoherent summaries. Besides, the performance becomes much worse when analyzing the per-

formance on role-oriented summaries and topic structures. We hope that this study could benchmark Chinese dialogue summarization and benefit further studies.

Back to Square One: Artifact Detection, Training and Commonsense Disentanglement in the Winograd Schema
Yanai Elazar et al.

The Winograd Schema (WS) has been proposed as a test for measuring commonsense capabilities of models. Recently, pre-trained language model-based approaches have boosted performance on some WS benchmarks but the source of improvement is still not clear. This paper suggests that the apparent progress on WS may not necessarily reflect progress in commonsense reasoning. To support this claim, we first show that the current evaluation method of WS is sub-optimal and propose a modification that uses twin sentences for evaluation. We also propose two new baselines that indicate the existence of artifacts in WS benchmarks. We then develop a method for evaluating WS-like sentences in a zero-shot setting to account for the commonsense reasoning abilities acquired during the pretraining and observe that popular language models perform randomly in this setting when using our more strict evaluation. We conclude that the observed progress is mostly due to the use of supervision in training WS models, which is not likely to successfully support all the required commonsense reasoning skills and knowledge.

Robustness Evaluation of Entity Disambiguation Using Prior Probes: the Case of Entity Overshadowing
Vera Provatorova et al.

Entity disambiguation (ED) is the last step of entity linking (EL), when candidate entities are reranked according to the context they appear in. All datasets for training and evaluating models for EL consist of convenience samples, such as news articles and tweets, that propagate the prior probability bias of the entity distribution towards more frequently occurring entities. It was shown that the performance of the EL systems on such datasets is overestimated since it is possible to obtain higher accuracy scores by merely learning the prior. To provide a more adequate evaluation benchmark, we introduce the ShadowLink dataset, which includes 16K short text snippets annotated with entity mentions. We evaluate and report the performance of popular EL systems on the ShadowLink benchmark. The results show a considerable difference in accuracy between more and less common entities for all of the EL systems under evaluation, demonstrating the effect of prior probability bias and entity overshadowing.

Building and Evaluating Open-Domain Dialogue Corpora with Clarifying Questions
Mohammad Ailannejadi et al.

Enabling open-domain dialogue systems to ask clarifying questions when appropriate is an important direction for improving the quality of the system response. Namely, for cases when a user request is not specific enough for a conversation system to provide an answer right away, it is desirable to ask a clarifying question to increase the chances of retrieving a satisfying answer. To address the problem of 'asking clarifying questions in open-domain dialogues': (1) we collect and release a new dataset focused on open-domain single- and multi-turn conversations, (2) we benchmark several state-of-the-art neural baselines, and (3) we propose a pipeline consisting of offline and online steps for evaluating the quality of clarifying questions in various dialogues. These contributions are suitable as a foundation for further research.

Agreeing to Disagree: Annotating Offensive Language Datasets with Annotators' Disagreement
Elisa Leonardelli et al.

Since state-of-the-art approaches to offensive language detection rely on supervised learning, it is crucial to quickly adapt them to the continuously evolving scenario of social media. While several approaches have been proposed to tackle the problem from an algorithmic perspective, so to reduce the need for annotated data, less attention has been paid to the quality of these data. Following a trend that has emerged recently, we focus on the level of agreement among annotators while selecting data to create offensive language datasets, a task involving a high level of subjectivity. Our study comprises the creation of three novel datasets of English tweets covering different topics and having five crowd-sourced judgments each. We also present an extensive set of experiments showing that selecting training and test data according to different levels of annotators' agreement has a strong effect on classifiers performance and robustness. Our findings are further validated in cross-domain experiments and studied using a popular benchmark dataset. We show that such hard cases, where low agreement is present, are not necessarily due to poor-quality annotation and we advocate for a higher presence of ambiguous cases in future datasets, in order to train more robust systems and better account for the different points of view expressed online.

PhoMT: A High-Quality and Large-Scale Benchmark Dataset for Vietnamese-English Machine Translation
Long Doan et al.

We introduce a high-quality and large-scale Vietnamese-English parallel dataset of 3.02M sentence pairs, which is 2.9M pairs larger than the benchmark Vietnamese-English machine translation corpus IWSLT15. We conduct experiments comparing strong neural baselines and well-known automatic translation engines on our dataset and find that in both automatic and human evaluations: the best performance is obtained by fine-tuning the pre-trained sequence-to-sequence denoising auto-encoder mBART. To our best knowledge, this is the first

large-scale Vietnamese-English machine translation study. We hope our publicly available dataset and study can serve as a starting point for future research and applications on Vietnamese-English machine translation. We release our dataset at: https://github.com/VinAIResearch/PhoMT

Lying Through One's Teeth: A Study on Verbal Leakage Cues

Min-Hsuan Yeh and Lun-Wei Ku

Although many studies use the LIWC lexicon to show the existence of verbal leakage cues in lie detection datasets, none mention how verbal leakage cues are influenced by means of data collection, or the impact thereof on the performance of models. In this paper, we study verbal leakage cues to understand the effect of the data construction method on their significance, and examine the relationship between such cues and models' validity. The LIWC word-category dominance scores of seven lie detection datasets are used to show that audio statements and lie-based annotations indicate a greater number of strong verbal leakage cue categories. Moreover, we evaluate the validity of state-of-the-art lie detection models with cross- and in-dataset testing. Results show that in both types of testing, models trained on a dataset with more strong verbal leakage cue categories—as opposed to only a greater number of strong cues—yield superior results, suggesting that verbal leakage cues are a key factor for selecting lie detection datasets.

Perturbation CheckLists for Evaluating NLG Evaluation Metrics

Ananya B. Sai et al.

Natural Language Generation (NLG) evaluation is a multifaceted task requiring assessment of multiple desirable criteria, e.g., fluency, coherency, coverage, relevance, adequacy, overall quality, etc. Across existing datasets for 6 NLG tasks, we observe that the human evaluation scores on these multiple criteria are often not correlated. For example, there is a very low correlation between human scores on fluency and data coverage for the task of structured data to text generation. This suggests that the current recipe of proposing new automatic evaluation metrics for NLG by showing that they correlate well with scores assigned by humans for a single criteria (overall quality) alone is inadequate. Indeed, our extensive study involving 25 automatic evaluation metrics across 6 different tasks and 18 different evaluation criteria shows that there is no single metric which correlates well with human scores on all desirable criteria, for most NLG tasks. Given this situation, we propose CheckLists for better design and evaluation of automatic metrics. We design templates which target a specific criteria (e.g., coverage) and perturb the output such that the quality gets affected only along this specific criteria (e.g., the coverage drops). We show that existing evaluation metrics are not robust against even such simple perturbations and disagree with scores assigned by humans to the perturbed output. The proposed templates thus allow for a fine-grained assessment of automatic evaluation metrics exposing their limitations and will facilitate better design, analysis and evaluation of such metrics. Our templates and code are available at https://iitmnlp.github.io/EvalEval/

Syntax Role for Neural Semantic Role Labeling

Zuchao Li et al.

Semantic role labeling (SRL) is dedicated to recognizing the semantic predicate-argument structure of a sentence. Previous studies in terms of traditional models have shown syntactic information can make remarkable contributions to SRL performance; however, the necessity of syntactic information was challenged by a few recent neural SRL studies that demonstrate impressive performance without syntactic backbones and suggest that syntax information becomes much less important for neural semantic role labeling, especially when paired with recent deep neural network and large-scale pre-trained language models. Despite this notion, the neural SRL field still lacks a systematic and full investigation on the relevance of syntactic information in SRL, for both dependency and both monolingual and multilingual settings. This paper intends to quantify the importance of syntactic information for neural SRL in the deep learning framework. We introduce three typical SRL frameworks (baselines), sequence-based, tree-based, and graph-based, which are accompanied by two categories of exploiting syntactic information: syntax pruning-based and syntax feature-based. Experiments are conducted on the CoNLL-2005, -2009, and -2012 benchmarks for all languages available, and results show that neural SRL models can still benefit from syntactic information under certain conditions. Furthermore, we show the quantitative significance of syntax to neural SRL models together with a thorough empirical survey using existing models.

Generating Datasets with Pretrained Language Models

Timo Schick and Hinrich Schütze

To obtain high-quality sentence embeddings from pretrained language models (PLMs), they must either be augmented with additional pretraining objectives or finetuned on a large set of labeled text pairs. While the latter approach typically outperforms the former, it requires great human effort to generate suitable datasets of sufficient size. In this paper, we show how PLMs can be leveraged to obtain high-quality sentence embeddings without the need for labeled data, finetuning or modifications to the pretraining objective: We utilize the generative abilities of large and high-performing PLMs to generate entire datasets of labeled text pairs from scratch, which we then use for finetuning much smaller and more efficient models. Our fully unsupervised approach outperforms strong baselines on several semantic textual similarity datasets.

Continuous Entailment Patterns for Lexical Inference in Context

Martin Schmitt and Hinrich Schütze

Combining a pretrained language model (PLM) with textual patterns has been shown to help in both zero- and few-shot settings. For zero-shot performance, it makes sense to design patterns that closely resemble

the text seen during self-supervised pretraining because the model has never seen anything else. Supervised training allows for more flexibility. If we allow for tokens outside the PLM's vocabulary, patterns can be adapted more flexibly to a PLM's idiosyncrasies. Contrasting patterns where a "token" can be any continuous vector from those where a discrete choice between vocabulary elements has to be made, we call our method CONtinous pAtterNs (CONAN). We evaluate CONAN on two established benchmarks for lexical inference in context (LIiC) a.k.a. predicate entailment, a challenging natural language understanding task with relatively small training data. In a direct comparison with discrete patterns, CONAN consistently leads to improved performance, setting a new state of the art. Our experiments give valuable insights on the kind of pattern that enhances a PLM's performance on LIiC and raise important questions regarding our understanding of PLMs using text patterns.

Fast, Effective, and Self-Supervised: Transforming Masked Language Models into Universal Lexical and Sentence Encoders
Fangyu Liu et al.

Previous work has indicated that pretrained Masked Language Models (MLMs) are not effective as universal lexical and sentence encoders off-the-shelf, i.e., without further task-specific fine-tuning on NLI, sentence similarity, or paraphrasing tasks using annotated task data. In this work, we demonstrate that it is possible to turn MLMs into effective lexical and sentence encoders even without any additional data, relying simply on self-supervision. We propose an extremely simple, fast, and effective contrastive learning technique, termed Mirror-BERT, which converts MLMs (e.g., BERT and RoBERTa) into such encoders in 20-30 seconds with no access to additional external knowledge. Mirror-BERT relies on identical and slightly modified string pairs as positive (i.e., synonymous) fine-tuning examples, and aims to maximise their similarity during "identity fine-tuning". We report huge gains over off-the-shelf MLMs with Mirror-BERT both in lexical-level and in sentence-level tasks, across different domains and different languages. Notably, in sentence similarity (STS) and question-answer entailment (QNLI) tasks, our self-supervised Mirror-BERT model even matches the performance of the Sentence-BERT models from prior work which rely on annotated task data. Finally, we delve deeper into the inner workings of MLMs, and suggest some evidence on why this simple Mirror-BERT fine-tuning approach can yield effective universal lexical and sentence encoders.

Multivalent Entailment Graphs for Question Answering
Nick McKenna et al.

Drawing inferences between open-domain natural language predicates is a necessity for true language understanding. There has been much progress in unsupervised learning of entailment graphs for this purpose. We make three contributions: (1) we reinterpret the Distributional Inclusion Hypothesis to model entailment between predicates of different valencies, like DEFEAT(Biden, Trump) entails WIN(Biden); (2) we actualize this theory by learning unsupervised Multivalent Entailment Graphs of open-domain predicates; and (3) we demonstrate the capabilities of these graphs on a novel question answering task. We show that directional entailment is more helpful for inference than non-directional similarity on questions of fine-grained semantics. We also show that drawing on evidence across valencies answers more questions than by using only the same valency evidence.

RuleBERT: Teaching Soft Rules to Pre-Trained Language Models
Mohammed Saeed et al.

While pre-trained language models (PLMs) are the go-to solution to tackle many natural language processing problems, they are still very limited in their ability to capture and to use common-sense knowledge. In fact, even if information is available in the form of approximate (soft) logical rules, it is not clear how to transfer it to a PLM in order to improve its performance for deductive reasoning tasks. Here, we aim to bridge this gap by teaching PLMs how to reason with soft Horn rules. We introduce a classification task where, given facts and soft rules, the PLM should return a prediction with a probability for a given hypothesis. We release the first dataset for this task, and we propose a revised loss function that enables the PLM to learn how to predict precise probabilities for the task. Our evaluation results show that the resulting fine-tuned models achieve very high performance, even on logical rules that were unseen at training. Moreover, we demonstrate that logical notions expressed by the rules are transferred to the fine-tuned model, yielding state-of-the-art results on external datasets.

VeeAlign: Multifaceted Context Representation Using Dual Attention for Ontology Alignment
Vivek Iyer, Arvind Agarwal, and Harshit Kumar

Ontology Alignment is an important research problem applied to various fields such as data integration, data transfer, data preparation, etc. State-of-the-art (SOTA) Ontology Alignment systems typically use naive domain-dependent approaches with handcrafted rules or domain-specific architectures, making them unscalable and inefficient. In this work, we propose VeeAlign, a Deep Learning based model that uses a novel dual-attention mechanism to compute the contextualized representation of a concept which, in turn, is used to discover alignments. By doing this, not only is our approach able to exploit both syntactic and semantic information encoded in ontologies, it is also, by design, flexible and scalable to different domains with minimal effort. We evaluate our model on four different datasets from different domains and languages, and establish its superiority through these results as well as detailed ablation studies. The code and datasets used are available at https://github.com/Remorax/VeeAlign.

Aligning Actions Across Recipe Graphs
Lucia Donatelli et al.
Recipe texts are an idiosyncratic form of instructional language that pose unique challenges for automatic understanding. One challenge is that a cooking step in one recipe can be explained in another recipe in different words, at a different level of abstraction, or not at all. Previous work has annotated correspondences between recipe instructions at the sentence level, often glossing over important correspondences between cooking steps across recipes. We present a novel and fully-parsed English recipe corpus, ARA (Aligned Recipe Actions), which annotates correspondences between individual actions across similar recipes with the goal of capturing information implicit for accurate recipe understanding. We represent this information in the form of recipe graphs, and we train a neural model for predicting correspondences on ARA. We find that substantial gains in accuracy can be obtained by taking fine-grained structural information about the recipes into account.

Stepmothers are mean and academics are pretentious: What do pretrained language models learn about you?
Rochelle Choenni, Ekaterina Shutova, and Robert van Rooij
In this paper, we investigate what types of stereotypical information are captured by pretrained language models. We present the first dataset comprising stereotypical attributes of a range of social groups and propose a method to elicit stereotypes encoded by pretrained language models in an unsupervised fashion. Moreover, we link the emergent stereotypes to their manifestation as basic emotions as a means to study their emotional effects in a more generalized manner. To demonstrate how our methods can be used to analyze emotion and stereotype shifts due to linguistic experience, we use fine-tuning on news sources as a case study. Our experiments expose how attitudes towards different social groups vary across models and how quickly emotions and stereotypes can shift at the fine-tuning stage.

Shortcutted Commonsense: Data Spuriousness in Deep Learning of Commonsense Reasoning
Ruben Branco et al.
Commonsense is a quintessential human capacity that has been a core challenge to Artificial Intelligence since its inception. Impressive results in Natural Language Processing tasks, including in commonsense reasoning, have consistently been achieved with Transformer neural language models, even matching or surpassing human performance in some benchmarks. Recently, some of these advances have been called into question: so called data artifacts in the training data have been made evident as spurious correlations and shallow shortcuts that in some cases are leveraging these outstanding results. In this paper we seek to further pursue this analysis into the realm of commonsense related language processing tasks. We undertake a study on different prominent benchmarks that involve commonsense reasoning, along a number of key stress experiments, thus seeking to gain insight on whether the models are learning transferable generalizations intrinsic to the problem at stake or just taking advantage of incidental shortcuts in the data items. The results obtained indicate that most datasets experimented with are problematic, with models resorting to non-robust features and appearing not to be learning and generalizing towards the overall tasks intended to be conveyed or exemplified by the datasets.

Finding needles in a haystack: Sampling Structurally-diverse Training Sets from Synthetic Data for Compositional Generalization
Inbar Oren, Jonathan Herzig, and Jonathan Berant
Modern semantic parsers suffer from two principal limitations. First, training requires expensive collection of utterance-program pairs. Second, semantic parsers fail to generalize at test time to new compositions/structures that have not been observed during training. Recent research has shown that automatic generation of synthetic utterance-program pairs can alleviate the first problem, but its potential for the second has thus far been underexplored. In this work, we investigate automatic generation of synthetic utterance-program pairs for improving compositional generalization in semantic parsing. Given a small training set of annotated examples and an "infinite" pool of synthetic examples, we select a subset of synthetic examples that are structurally-diverse and use them to improve compositional generalization. We evaluate our approach on a new split of the schema2QA dataset, and show that it leads to dramatic improvements in compositional generalization as well as moderate improvements in the traditional i.i.d setup. Moreover, structurally-diverse sampling achieves these improvements with as few as 5K examples, compared to 1M examples when sampling uniformly at random – a 200x improvement in data efficiency.

GeneSis: A Generative Approach to Substitutes in Context
Caterina Lacerra, Rocco Tripodi, and Roberto Navigli
The lexical substitution task aims at generating a list of suitable replacements for a target word in context, ideally keeping the meaning of the modified text unchanged. While its usage has increased in recent years, the paucity of annotated data prevents the finetuning of neural models on the task, hindering the full fruition of recently introduced powerful architectures such as language models. Furthermore, lexical substitution is usually evaluated in a framework that is strictly bound to a limited vocabulary, making it impossible to credit appropriate, but out-of-vocabulary, substitutes. To assess these issues, we proposed GeneSis (Generating Substitutes in contexts), the first generative approach to lexical substitution. Thanks to a seq2seq model, we generate substitutes for a word according to the context it appears in, attaining state-of-the-art results on different benchmarks. Moreover, our approach allows silver data to be produced for further improving the performances of lexical substitution systems. Along with an extensive analysis of GeneSis results, we also present a human evaluation of the generated substitutes in order to assess their quality. We release the fine-tuned models, the generated

datasets, and the code to reproduce the experiments at https://github.com/SapienzaNLP/genesis.

Exploiting Twitter as Source of Large Corpora of Weakly Similar Pairs for Semantic Sentence Embeddings
Marco Di Giovanni and Marco Brambilla

Semantic sentence embeddings are usually supervisedly built minimizing distances between pairs of embeddings of sentences labelled as semantically similar by annotators. Since big labelled datasets are rare, in particular for non-English languages, and expensive, recent studies focus on unsupervised approaches that require not-paired input sentences. We instead propose a language-independent approach to build large datasets of pairs of informal texts weakly similar, without manual human effort, exploiting Twitter's intrinsic powerful signals of relatedness: replies and quotes of tweets. We use the collected pairs to train a Transformer model with triplet-like structures, and we test the generated embeddings on Twitter NLP similarity tasks (PIT and TURL) and STSb. We also introduce four new sentence ranking evaluation benchmarks of informal texts, carefully extracted from the initial collections of tweets, proving not only that our best model learns classical Semantic Textual Similarity, but also excels on tasks where pairs of sentences are not exact paraphrases. Ablation studies reveal how increasing the corpus size influences positively the results, even at 2M samples, suggesting that bigger collections of Tweets still do not contain redundant information about semantic similarities. Code available at https://github.com/marco-digio/Twitter4SSE

Asking It All: Generating Contextualized Questions for any Semantic Role
Valentina Pyatkin et al.

Asking questions about a situation is an inherent step towards understanding it. To this end, we introduce the task of role question generation, which, given a predicate mention and a passage, requires producing a set of questions asking about all possible semantic roles of the predicate. We develop a two-stage model for this task, which first produces a context-independent question prototype for each role and then revises it to be contextually appropriate for the passage. Unlike most existing approaches to question generation, our approach does not require conditioning on existing answers in the text. Instead, we condition on the type of information to inquire about, regardless of whether the answer appears explicitly in the text, could be inferred from it, or should be sought elsewhere. Our evaluation demonstrates that we generate diverse and well-formed questions for a large, broad-coverage ontology of predicates and roles.

ConSeC: Word Sense Disambiguation as Continuous Sense Comprehension
Edoardo Barba, Luigi Procopio, and Roberto Navigli

Supervised systems have nowadays become the standard recipe for Word Sense Disambiguation (WSD), with Transformer-based language models as their primary ingredient. However, while these systems have certainly attained unprecedented performances, virtually all of them operate under the constraining assumption that, given a context, each word can be disambiguated individually with no account of the other sense choices. To address this limitation and drop this assumption, we propose CONtinuous SEnse Comprehension (ConSeC), a novel approach to WSD: leveraging a recent re-framing of this task as a text extraction problem, we adapt it to our formulation and introduce a feedback loop strategy that allows the disambiguation of a target word to be conditioned not only on its context but also on the explicit senses assigned to nearby words. We evaluate ConSeC and examine how its components lead it to surpass all its competitors and set a new state of the art on English WSD. We also explore how ConSeC fares in the cross-lingual setting, focusing on 8 languages with various degrees of resource availability, and report significant improvements over prior systems. We release our code at https://github.com/SapienzaNLP/consec.

Semi-Supervised Exaggeration Detection of Health Science Press Releases
Dustin Wright and Isabelle Augenstein

Public trust in science depends on honest and factual communication of scientific papers. However, recent studies have demonstrated a tendency of news media to misrepresent scientific papers by exaggerating their findings. Given this, we present a formalization of and study into the problem of exaggeration detection in science communication. While there are an abundance of scientific papers and popular media articles written about them, very rarely do the articles include a direct link to the original paper, making data collection challenging, and necessitating the need for few-shot learning. We address this by curating a set of labeled press release/abstract pairs from existing expert annotated studies on exaggeration in press releases of scientific papers suitable for benchmarking the performance of machine learning models on the task. Using limited data from this and previous studies on exaggeration detection in science, we introduce MT-PET, a multi-task version of Pattern Exploiting Training (PET), which leverages knowledge from complementary cloze-style QA tasks to improve few-shot learning. We demonstrate that MT-PET outperforms PET and supervised learning both when data is limited, as well as when there is an abundance of data for the main task.

QA-Align: Representing Cross-Text Content Overlap by Aligning Question-Answer Propositions
Daniela Brook Weiss et al.

Multi-text applications, such as multi-document summarization, are typically required to model redundancies across related texts. Current methods confronting consolidation struggle to fuse overlapping information. In order to explicitly represent content overlap, we propose to align predicate-argument relations across texts, providing a potential scaffold for information consolidation. We go beyond clustering coreferring mentions,

and instead model overlap with respect to redundancy at a propositional level, rather than merely detecting shared referents. Our setting exploits QA-SRL, utilizing question-answer pairs to capture predicate-argument relations, facilitating laymen annotation of cross-text alignments. We employ crowd-workers for constructing a dataset of QA-based alignments, and present a baseline QA alignment model trained over our dataset. Analyses show that our new task is semantically challenging, capturing content overlap beyond lexical similarity and complements cross-document coreference with proposition-level links, offering potential use for downstream tasks.

Detecting Contact-Induced Semantic Shifts: What Can Embedding-Based Methods Do in Practice?
Filip Miletic, Anne Przewozny-Desriaux, and Ludovic Tanguy

This study investigates the applicability of semantic change detection methods in descriptively oriented linguistic research. It specifically focuses on contact-induced semantic shifts in Quebec English. We contrast synchronic data from different regions in order to identify the meanings that are specific to Quebec and potentially related to language contact. Type-level embeddings are used to detect new semantic shifts, and token-level embeddings to isolate regionally specific occurrences. We introduce a new 80-item test set and conduct both quantitative and qualitative evaluations. We demonstrate that diachronic word embedding methods can be applied to contact-induced semantic shifts observed in synchrony, obtaining results comparable to the state of the art on similar tasks in diachrony. However, we show that encouraging evaluation results do not translate to practical value in detecting new semantic shifts. Finally, our application of token-level embeddings accelerates manual data exploration and provides an efficient way of scaling up sociolinguistic analyses.

Guilt by Association: Emotion Intensities in Lexical Representations
Shahab Raji and Gerard de Melo

What do linguistic models reveal about the emotions associated with words? In this study, we consider the task of estimating word-level emotion intensity scores for specific emotions, exploring unsupervised, supervised, and finally a self-supervised method of extracting emotional associations from pretrained vectors and models. Overall, we find that linguistic models carry substantial potential for inducing fine-grained emotion intensity scores, showing a far higher correlation with human ground truth ratings than state-of-the-art emotion lexicons based on labeled data.

Beta Distribution Guided Aspect-aware Graph for Aspect Category Sentiment Analysis with Affective Knowledge
Bin Liang et al.

In this paper, we investigate the Aspect Category Sentiment Analysis (ACSA) task from a novel perspective by exploring a Beta Distribution guided aspect-aware graph construction based on external knowledge. That is, we are no longer entangled about how to laboriously search the sentiment clues for coarse-grained aspects from the context, but how to preferably find the words highly related to the aspects in the context and determine their importance based on the public knowledge base. In this way, the contextual sentiment clues can be explicitly tracked in ACSA for the aspects in the light of these aspect-related words. To be specific, we first regard each aspect as a pivot to derive aspect-aware words that are highly related to the aspect from external affective commonsense knowledge. Then, we employ Beta Distribution to educe the aspect-aware weight, which reflects the importance to the aspect, for each aspect-aware word. Afterward, the aspect-aware words are served as the substitutes of the coarse-grained aspect to construct graphs for leveraging the aspect-related contextual sentiment dependencies in ACSA. Experiments on 6 benchmark datasets show that our approach significantly outperforms the state-of-the-art baseline methods.

DILBERT: Customized Pre-Training for Domain Adaptation with Category Shift, with an Application to Aspect Extraction
Entony Lekhtman, Yftah Ziser, and Roi Reichart

The rise of pre-trained language models has yielded substantial progress in the vast majority of Natural Language Processing (NLP) tasks. However, a generic approach towards the pre-training procedure can naturally be sub-optimal in some cases. Particularly, fine-tuning a pre-trained language model on a source domain and then applying it to a different target domain, results in a sharp performance decline of the eventual classifier for many source-target domain pairs. Moreover, in some NLP tasks, the output categories substantially differ between domains, making adaptation even more challenging. This, for example, happens in the task of aspect extraction, where the aspects of interest of reviews of, e.g., restaurants or electronic devices may be very different. This paper presents a new fine-tuning scheme for BERT, which aims to address the above challenges. We name this scheme DILBERT: Domain Invariant Learning with BERT, and customize it for aspect extraction in the unsupervised domain adaptation setting. DILBERT harnesses the categorical information of both the source and the target domains to guide the pre-training process towards a more domain and category invariant representation, thus closing the gap between the domains. We show that DILBERT yields substantial improvements over state-of-the-art baselines while using a fraction of the unlabeled data, particularly in more challenging domain adaptation setups.

On Classifying whether Two Texts are on the Same Side of an Argument
Erik Körner et al.

To ease the difficulty of argument stance classification, the task of same side stance classification (S3C) has been proposed. In contrast to actual stance classification, which requires a substantial amount of domain knowledge to identify whether an argument is in favor or against a certain issue, it is argued that, for S3C, only argument similarity within stances needs to be learned to successfully solve the task. We evaluate several transformer-based approaches on the dataset of the recent S3C shared task, followed by an in-depth evaluation and error analysis of our model and the task's hypothesis. We show that, although we achieve state-of-the-art results, our model fails to generalize both within as well as across topics and domains when adjusting the sampling strategy of the training and test set to a more adversarial scenario. Our evaluation shows that current state-of-the-art approaches cannot determine same side stance by considering only domain-independent linguistic similarity features, but appear to require domain knowledge and semantic inference, too.

Improving Multimodal fusion via Mutual Dependency Maximisation
Pierre Colombo et al.

Multimodal sentiment analysis is a trending area of research, and multimodal fusion is one of its most active topic. Acknowledging humans communicate through a variety of channels (i.e visual, acoustic, linguistic), multimodal systems aim at integrating different unimodal representations into a synthetic one. So far, a consequent effort has been made on developing complex architectures allowing the fusion of these modalities. However, such systems are mainly trained by minimising simple losses such as L_1 or cross-entropy. In this work, we investigate unexplored penalties and propose a set of new objectives that measure the dependency between modalities. We demonstrate that our new penalties lead to a consistent improvement (up to 4.3 on accuracy) across a large variety of state-of-the-art models on two well-known sentiment analysis datasets: `CMU-MOSI` and `CMU-MOSEI`. Our method not only achieves a new SOTA on both datasets but also produces representations that are more robust to modality drops. Finally, a by-product of our methods includes a statistical network which can be used to interpret the high dimensional representations learnt by the model.

The Effect of Round-Trip Translation on Fairness in Sentiment Analysis
Jonathan Christiansen, Mathias Gammelgaard, and Anders Søgaard

Sentiment analysis systems have been shown to exhibit sensitivity to protected attributes. Round-trip translation, on the other hand, has been shown to normalize text. We explore the impact of round-trip translation on the demographic parity of sentiment classifiers and show how round-trip translation consistently improves classification fairness at test time (reducing up to 47% of between-group gaps). We also explore the idea of retraining sentiment classifiers on round-trip-translated data.

Tribrid: Stance Classification with Neural Inconsistency Detection
Song Yang and Jacopo Urbani

We study the problem of performing automatic stance classification on social media with neural architectures such as BERT. Although these architectures deliver impressive results, their level is not yet comparable to the one of humans and they might produce errors that have a significant impact on the downstream task (e.g., fact-checking). To improve the performance, we present a new neural architecture where the input also includes automatically generated negated perspectives over a given claim. The model is jointly learned to make simultaneously multiple predictions, which can be used either to improve the classification of the original perspective or to filter out doubtful predictions. In the first case, we propose a weakly supervised method for combining the predictions into a final one. In the second case, we show that using the confidence scores to remove doubtful predictions allows our method to achieve human-like performance over the retained information, which is still a sizable part of the original input.

Few-Shot Emotion Recognition in Conversation with Sequential Prototypical Networks
Gaël Guibon et al.

Several recent studies on dyadic human-human interactions have been done on conversations without specific business objectives. However, many companies might benefit from studies dedicated to more precise environments such as after sales services or customer satisfaction surveys. In this work, we place ourselves in the scope of a live chat customer service in which we want to detect emotions and their evolution in the conversation flow. This context leads to multiple challenges that range from exploiting restricted, small and mostly unlabeled datasets to finding and adapting methods for such context. We tackle these challenges by using Few-Shot Learning while making the hypothesis it can serve conversational emotion classification for different languages and sparse labels. We contribute by proposing a variation of Prototypical Networks for sequence labeling in conversation that we name ProtoSeq. We test this method on two datasets with different languages: daily conversations in English and customer service chat conversations in French. When applied to emotion classification in conversations, our method proved to be competitive even when compared to other ones.

Not All Negatives are Equal: Label-Aware Contrastive Loss for Fine-grained Text Classification
Varsha Suresh and Desmond Ong

Fine-grained classification involves dealing with datasets with larger number of classes with subtle differences between them. Guiding the model to focus on differentiating dimensions between these commonly confusable classes is key to improving performance on fine-grained tasks. In this work, we analyse the contrastive fine-tuning of pre-trained language models on two fine-grained text classification tasks, emotion classification and

sentiment analysis. We adaptively embed class relationships into a contrastive objective function to help differently weigh the positives and negatives, and in particular, weighting closely confusable negatives more than less similar negative examples. We find that Label-aware Contrastive Loss outperforms previous contrastive methods, in the presence of larger number and/or more confusable classes, and helps models to produce output distributions that are more differentiated.

Inflate and Shrink:Enriching and Reducing Interactions for Fast Text-Image Retrieval

Haoliang Liu, Tan Yu, and Ping Li

By exploiting the cross-modal attention, cross-BERT methods have achieved state-of-the-art accuracy in cross-modal retrieval. Nevertheless, the heavy text-image interactions in the cross-BERT model are prohibitively slow for large-scale retrieval. Late-interaction methods trade off retrieval accuracy and efficiency by exploiting cross-modal interaction only in the late stage, attaining a satisfactory retrieval speed. In this work, we propose an inflating and shrinking approach to further boost the efficiency and accuracy of late-interaction methods. The inflating operation plugs several codes in the input of the encoder to exploit the text-image interactions more thoroughly for higher retrieval accuracy. Then the shrinking operation gradually reduces the text-image interactions through knowledge distilling for higher efficiency. Through an inflating operation followed by a shrinking operation, both efficiency and accuracy of a late-interaction model are boosted. Systematic experiments on public benchmarks demonstrate the effectiveness of our inflating and shrinking approach.

COVR: A Test-Bed for Visually Grounded Compositional Generalization with Real Images

Ben Bogin et al.

While interest in models that generalize at test time to new compositions has risen in recent years, benchmarks in the visually-grounded domain have thus far been restricted to synthetic images. In this work, we propose COVR, a new test-bed for visually-grounded compositional generalization with real images. To create COVR, we use real images annotated with scene graphs, and propose an almost fully automatic procedure for generating question-answer pairs along with a set of context images. COVR focuses on questions that require complex reasoning, including higher-order operations such as quantification and aggregation. Due to the automatic generation process, COVR facilitates the creation of compositional splits, where models at test time need to generalize to new concepts and compositions in a zero- or few-shot setting. We construct compositional splits using COVR and demonstrate a myriad of cases where state-of-the-art pre-trained language-and-vision models struggle to compositionally generalize.

Learning grounded word meaning representations on similarity graphs

Mariella Dimiccoli, Herwig Wendt, and Pau Batlle Franch

This paper introduces a novel approach to learn visually grounded meaning representations of words as low-dimensional node embeddings on an underlying graph hierarchy. The lower level of the hierarchy models modality-specific word representations, conditioned to another modality, through dedicated but communicating graphs, while the higher level puts these representations together on a single graph to learn a representation jointly from both modalities. The topology of each graph models similarity relations among words, and is estimated jointly with the graph embedding. The assumption underlying this model is that words sharing similar meaning correspond to communities in an underlying graph in a low-dimensional space. We named this model Hierarchical Multi-Modal Similarity Graph Embedding (HM-SGE). Experimental results validate the ability of HM-SGE to simulate human similarity judgments and concept categorization, outperforming the state of the art.

Vision-and-Language or Vision-for-Language? On Cross-Modal Influence in Multimodal Transformers

Stella Frank, Emanuele Bugliarello, and Desmond Elliott

Pretrained vision-and-language BERTs aim to learn representations that combine information from both modalities. We propose a diagnostic method based on cross-modal input ablation to assess the extent to which these models actually integrate cross-modal information. This method involves ablating inputs from one modality, either entirely or selectively based on cross-modal grounding alignments, and evaluating the model prediction performance on the other modality. Model performance is measured by modality-specific tasks that mirror the model pretraining objectives (e.g. masked language modelling for text). Models that have learned to construct cross-modal representations using both modalities are expected to perform worse when inputs are missing from a modality. We find that recently proposed models have much greater relative difficulty predicting text when visual information is ablated, compared to predicting visual object categories when text is ablated, indicating that these models are not symmetrically cross-modal.

HypMix: Hyperbolic Interpolative Data Augmentation

Ramit Sawhney et al.

Interpolation-based regularisation methods for data augmentation have proven to be effective for various tasks and modalities. These methods involve performing mathematical operations over the raw input samples or their latent states representations - vectors that often possess complex hierarchical geometries. However, these operations are performed in the Euclidean space, simplifying these representations, which may lead to distorted and noisy interpolations. We propose HypMix, a novel model-, data-, and modality-agnostic interpolative data augmentation technique operating in the hyperbolic space, which captures the complex geometry of input and

hidden state hierarchies better than its contemporaries. We evaluate HypMix on benchmark and low resource datasets across speech, text, and vision modalities, showing that HypMix consistently outperforms state-of-the-art data augmentation techniques. In addition, we demonstrate the use of HypMix in semi-supervised settings. We further probe into the adversarial robustness and qualitative inferences we draw from HypMix that elucidate the efficacy of the Riemannian hyperbolic manifolds for interpolation-based data augmentation.

Aspect-Controllable Opinion Summarization
Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata

Recent work on opinion summarization produces general summaries based on a set of input reviews and the popularity of opinions expressed in them. In this paper, we propose an approach that allows the generation of customized summaries based on aspect queries (e.g., describing the location and room of a hotel). Using a review corpus, we create a synthetic training dataset of (review, summary) pairs enriched with aspect controllers which are induced by a multi-instance learning model that predicts the aspects of a document at different levels of granularity. We fine-tune a pretrained model using our synthetic dataset and generate aspect-specific summaries by modifying the aspect controllers. Experiments on two benchmarks show that our model outperforms the previous state of the art and generates personalized summaries by controlling the number of aspects discussed in them.

QuestEval: Summarization Asks for Fact-based Evaluation
Thomas Scialom et al.

Summarization evaluation remains an open research problem: current metrics such as ROUGE are known to be limited and to correlate poorly with human judgments. To alleviate this issue, recent work has proposed evaluation metrics which rely on question answering models to assess whether a summary contains all the relevant information in its source document. Though promising, the proposed approaches have so far failed to correlate better than ROUGE with human judgments. In this paper, we extend previous approaches and propose a unified framework, named QuestEval. In contrast to established metrics such as ROUGE or BERTScore, QuestEval does not require any ground-truth reference. Nonetheless, QuestEval substantially improves the correlation with human judgments over four evaluation dimensions (consistency, coherence, fluency, and relevance), as shown in extensive experiments.

MassiveSumm: a very large-scale, very multilingual, news summarisation dataset
Daniel Varab and Natalie Schluter

Current research in automatic summarisation is unapologetically anglo-centered–a persistent state-of-affairs, which also predates neural net approaches. High-quality automatic summarisation datasets are notoriously expensive to create, posing a challenge for any language. However, with digitalisation, archiving, and social media advertising of newswire articles, recent work has shown how, with careful methodology application, large-scale datasets can now be simply gathered instead of written. In this paper, we present a large-scale multilingual summarisation dataset containing articles in 92 languages, spread across 28.8 million articles, in more than 35 writing scripts. This is both the largest, most inclusive, existing automatic summarisation dataset, as well as one of the largest, most inclusive, ever published datasets for any NLP task. We present the first investigation on the efficacy of resource building from news platforms in the low-resource language setting. Finally, we provide some first insight on how low-resource language settings impact state-of-the-art automatic summarisation system performance.

A Thorough Evaluation of Task-Specific Pretraining for Summarization
Sascha Rothe, Joshua Maynez, and Shashi Narayan

Task-agnostic pretraining objectives like masked language models or corrupted span prediction are applicable to a wide range of NLP downstream tasks (Raffel et al.,2019), but are outperformed by task-specific pretraining objectives like predicting extracted gap sentences on summarization (Zhang et al.,2020). We compare three summarization specific pretraining objectives with the task agnostic corrupted span prediction pretraining in controlled study. We also extend our study to a low resource and zero shot setup, to understand how many training examples are needed in order to ablate the task-specific pretraining without quality loss. Our results show that task-agnostic pretraining is sufficient for most cases which hopefully reduces the need for costly task-specific pretraining. We also report new state-of-the-art number for two summarization task using a T5 model with 11 billion parameters and an optimal beam search length penalty.

Controllable Summarization with Constrained Markov Decision Process
Hou Pong Chan, Lu Wang, and Irwin King

We study controllable text summarization which allows users to gain control on a particular attribute (e.g., length limit) of the generated summaries. In this work, we propose a novel training framework based on Constrained Markov Decision Process (CMDP), which conveniently includes a reward function along with a set of constraints, to facilitate better summarization control. The reward function encourages the generation to resemble the human-written reference, while the constraints are used to explicitly prevent the generated summaries from violating user-imposed requirements. Our framework can be applied to control important attributes of summarization, including length, covered entities, and abstractiveness, as we devise specific constraints for each of these aspects. Extensive experiments on popular benchmarks show that our CMDP framework helps generate informative summaries while complying with a given attribute's requirement.

Reducing Discontinuous to Continuous Parsing with Pointer Network Reordering
Daniel Fernández-González and Carlos Gómez-Rodríguez

Discontinuous constituent parsers have always lagged behind continuous approaches in terms of accuracy and speed, as the presence of constituents with discontinuous yield introduces extra complexity to the task. However, a discontinuous tree can be converted into a continuous variant by reordering tokens. Based on that, we propose to reduce discontinuous parsing to a continuous problem, which can then be directly solved by any off-the-shelf continuous parser. To that end, we develop a Pointer Network capable of accurately generating the continuous token arrangement for a given input sentence and define a bijective function to recover the original order. Experiments on the main benchmarks with two continuous parsers prove that our approach is on par in accuracy with purely discontinuous state-of-the-art algorithms, but considerably faster.

On the Relation between Syntactic Divergence and Zero-Shot Performance
Ofir Arviv et al.

We explore the link between the extent to which syntactic relations are preserved in translation and the ease of correctly constructing a parse tree in a zero-shot setting. While previous work suggests such a relation, it tends to focus on the macro level and not on the level of individual edges—a gap we aim to address. As a test case, we take the transfer of Universal Dependencies (UD) parsing from English to a diverse set of languages and conduct two sets of experiments. In one, we analyze zero-shot performance based on the extent to which English source edges are preserved in translation. In another, we apply three linguistically motivated transformations to UD, creating more cross-lingually stable versions of it, and assess their zero-shot parsability. In order to compare parsing performance across different schemes, we perform extrinsic evaluation on the downstream task of cross-lingual relation extraction (RE) using a subset of a standard English RE benchmark translated to Russian and Korean. In both sets of experiments, our results suggest a strong relation between cross-lingual stability and zero-shot parsing performance.

A Root of a Problem: Optimizing Single-Root Dependency Parsing
Miloš Stanojević and Shay B. Cohen

We describe two approaches to single-root dependency parsing that yield significant speed ups in such parsing. One approach has been previously used in dependency parsers in practice, but remains undocumented in the parsing literature, and is considered a heuristic. We show that this approach actually finds the optimal dependency tree. The second approach relies on simple reweighting of the inference graph being input to the dependency parser and has an optimal running time. Here, we again show that this approach is fully correct and identifies the highest-scoring parse tree. Our experiments demonstrate a manyfold speed up compared to a previous graph-based state-of-the-art parser without any loss in accuracy or optimality.

Efficient Sampling of Dependency Structure
Ran Zmigrod, Tim Vieira, and Ryan Cotterell

Probabilistic distributions over spanning trees in directed graphs are a fundamental model of dependency structure in natural language processing, syntactic dependency trees. In NLP, dependency trees often have an additional root constraint: only one edge may emanate from the root. However, no sampling algorithm has been presented in the literature to account for this additional constraint. In this paper, we adapt two spanning tree sampling algorithms to faithfully sample dependency trees from a graph subject to the root constraint. Wilson (1996('s sampling algorithm has a running time of O(H) where H is the mean hitting time of the graph. Colbourn (1996)'s sampling algorithm has a running time of $O(N^3)$, which is often greater than the mean hitting time of a directed graph. Additionally, we build upon Colbourn's algorithm and present a novel extension that can sample K trees without replacement in $O(K\ N^3 + K^2\ N)$ time. To the best of our knowledge, no algorithm has been given for sampling spanning trees without replacement from a directed graph.

Virtual Demo Session

MiSS: An Assistant for Multi-Style Simultaneous Translation
Zuchao Li et al.

In this paper, we present **MiSS**, an assistant for multi-style simultaneous translation. Our proposed translation system has five key features: highly accurate translation, simultaneous translation, translation for multiple text styles, back-translation for translation quality evaluation, and grammatical error correction. With this system, we aim to provide a complete translation experience for machine translation users. Our design goals are high translation accuracy, real-time translation, flexibility, and measurable translation quality. Compared with the free commercial translation systems commonly used, our translation assistance system regards the machine translation application as a more complete and fully-featured tool for users. By incorporating additional features and giving the user better control over their experience, we improve translation efficiency and performance. Additionally, our assistant system combines machine translation, grammatical error correction, and interactive edits, and uses a crowdsourcing mode to collect more data for further training to improve both the machine translation and grammatical error correction models. A short video demonstrating our system is available at https://www.youtube.com/watch?v=ZGCo7KtRKd8.

TransIns: Document Translation with Markup Reinsertion
Jörg Steffen and Josef van Genabithi

For many use cases, it is required that MT does not just translate raw text, but complex formatted documents (e.g. websites, slides, spreadsheets) and the result of the translation should reflect the formatting. This is challenging, as markup can be nested, apply to spans contiguous in source but non-contiguous in target etc. Here we present TransIns, a system for non-plain text document translation that builds on the Okapi framework and MT models trained with Marian NMT. We develop, implement and evaluate different strategies for reinserting markup into translated sentences using token alignments between source and target sentences. We propose a simple and effective strategy that compiles down all markup to single source tokens and transfers them to aligned target tokens. A first evaluation shows that this strategy yields highly accurate markup in the translated documents that outperforms the markup quality found in documents translated with popular translation services. We release TransIns under the MIT License as open-source software on https://github.com/DFKI-MLT/TransIns. An online demonstrator is available at https://transins.dfki.de.

KOAS: Korean Text Offensiveness Analysis System

San-Hee Park et al.

Warning: This manuscript contains a certain level of offensive expression. As communication through social media platforms has grown immensely, the increasing prevalence of offensive language online has become a critical problem. Notably in Korea, one of the countries with the highest Internet usage, automatic detection of offensive expressions has recently been brought to attention. However, morphological richness and complex syntax of Korean causes difficulties in neural model training. Furthermore, most of previous studies mainly focus on the detection of abusive language, disregarding implicit offensiveness and underestimating a different degree of intensity. To tackle these problems, we present KOAS, a system that fully exploits both contextual and linguistic features and estimates an offensiveness score for a text. We carefully designed KOAS with a multi-task learning framework and constructed a Korean dataset for offensive analysis from various domains. Refer for a detailed demonstration.

fairseq S^2: A Scalable and Integrable Speech Synthesis Toolkit

Changhan Wang et al.

This paper presents fairseq S^2, a fairseq extension for speech synthesis. We implement a number of autoregressive (AR) and non-AR text-to-speech models, and their multi-speaker variants. To enable training speech synthesis models with less curated data, a number of preprocessing tools are built and their importance is shown empirically. To facilitate faster iteration of development and analysis, a suite of automatic metrics is included. Apart from the features added specifically for this extension, fairseq S^2 also benefits from the scalability offered by fairseq and can be easily integrated with other state-of-the-art systems provided in this framework. The code, documentation, and pre-trained models will be made available at https://github.com/pytorch/fairseq/tree/master/examples/speech$_s ynthesis$.

UMR-Writer: A Web Application for Annotating Uniform Meaning Representations

Jin Zhao et al.

We present UMR-Writer, a web-based application for annotating Uniform Meaning Representations (UMR), a graph-based, cross-linguistically applicable semantic representation developed recently to support the development of interpretable natural language applications that require deep semantic analysis of texts. We present the functionalities of UMR-Writer and discuss the challenges in developing such a tool and how they are addressed.

Summary Explorer: Visualizing the State of the Art in Text Summarization

Shahbaz Syed et al.

This paper introduces Summary Explorer, a new tool to support the manual inspection of text summarization systems by compiling the outputs of 55 state-of-the-art single document summarization approaches on three benchmark datasets, and visually exploring them during a qualitative assessment. The underlying design of the tool considers three well-known summary quality criteria (coverage, faithfulness, and position bias), encapsulated in a guided assessment based on tailored visualizations. The tool complements existing approaches for locally debugging summarization models and improves upon them. The tool is available at https://tldr.webis.de/

Box Embeddings: An open-source library for representation learning using geometric structures

Tejas Chheda et al.

A fundamental component to the success of modern representation learning is the ease of performing various vector operations. Recently, objects with more geometric structure (eg. distributions, complex or hyperbolic vectors, or regions such as cones, disks, or boxes) have been explored for their alternative inductive biases and additional representational capacity. In this work, we introduce Box Embeddings, a Python library that enables researchers to easily apply and extend probabilistic box embeddings. Fundamental geometric operations on boxes are implemented in a numerically stable way, as are modern approaches to training boxes which mitigate gradient sparsity. The library is fully open source, and compatible with both PyTorch and TensorFlow, which allows existing neural network layers to be replaced with or transformed into boxes easily. In this work, we present the implementation details of the fundamental components of the library, and the concepts required to use box representations alongside existing neural network architectures.

LexiClean: An annotation tool for rapid multi-task lexical normalisation
Tyler Bikaun et al.

NLP systems are often challenged by difficulties arising from noisy, non-standard, and domain specific corpora. The task of lexical normalisation aims to standardise such corpora, but currently lacks suitable tools to acquire high-quality annotated data to support deep learning based approaches. In this paper, we present LexiClean, the first open-source web-based annotation tool for multi-task lexical normalisation. LexiClean's main contribution is support for simultaneous in situ token-level modification and annotation that can be rapidly applied corpus wide. We demonstrate the usefulness of our tool through a case study on two sets of noisy corpora derived from the specialised-domain of industrial mining. We show that LexiClean allows for the rapid and efficient development of high-quality parallel corpora. A demo of our system is available at: https://youtu.be/P7$_{o}oKrQPDU$.

OpenFraming: Open-sourced Tool for Computational Framing Analysis of Multilingual Data
Vibhu Bhatia et al.

When journalists cover a news story, they can cover the story from multiple angles or perspectives. These perspectives are called "frames" and usage of one frame or another may influence public perception and opinion of the issue at hand. We develop a web-based system for analyzing frames in multilingual text documents. We propose and guide users through a five-step end-to-end computational framing analysis framework grounded in media framing theory in communication research. Users can use the framework to analyze multilingual text data, starting from the exploration of frames in user corpora and through review of previous framing literature (step 1-3) to frame classification (step 4) and prediction (step 5). The framework combines unsupervised and supervised machine learning and leverages a state-of-the-art (SoTA) multilingual language model, which can significantly enhance frame prediction performance while requiring a considerably small sample of manual annotations. Through the interactive website, anyone can perform the proposed computational framing analysis, making advanced computational analysis available to researchers without a programming background and bridging the digital divide within the communication research discipline in particular and the academic community in general. The system is available online at http://www.openframing.org, via an API http://www.openframing.org:5000/docs/, or through our GitHub page https://github.com/vibss2397/openFraming.

CroAno : A Crowd Annotation Platform for Improving Label Consistency of Chinese NER Dataset
Baoli Zhang et al.

In this paper, we introduce CroAno, a web-based crowd annotation platform for the Chinese named entity recognition (NER). Besides some basic features for crowd annotation like fast tagging and data management, CroAno provides a systematic solution for improving label consistency of Chinese NER dataset. 1) Disagreement Adjudicator: CroAno uses a multi-dimensional highlight mode to visualize instance-level inconsistent entities and makes the revision process user-friendly. 2) Inconsistency Detector: CroAno employs a detector to locate corpus-level label inconsistency and provides users an interface to correct inconsistent entities in batches. 3) Prediction Error Analyzer: We deconstruct the entity prediction error of the model to six fine-grained entity error types. Users can employ this error system to detect corpus-level inconsistency from a model perspective. To validate the effectiveness of our platform, we use CroAno to revise two public datasets. In the two revised datasets, we get an improvement of +1.96% and +2.57% F1 respectively in model performance.

iFacetSum: Coreference-based Interactive Faceted Summarization for Multi-Document Exploration
Eran Hirsch et al.

We introduce iFacetSum, a web application for exploring topical document collections. iFacetSum integrates interactive summarization together with faceted search, by providing a novel faceted navigation scheme that yields abstractive summaries for the user selections. This approach offers both a comprehensive overview as well as particular details regarding subtopics of choice. The facets are automatically produced based on cross-document coreference pipelines, rendering generic concepts, entities and statements surfacing in the source texts. We analyze the effectiveness of our application through small-scale user studies that suggest the usefulness of our tool.

AMuSE-WSD: An All-in-one Multilingual System for Easy Word Sense Disambiguation
Riccardo Orlando et al.

Over the past few years, Word Sense Disambiguation (WSD) has received renewed interest: recently proposed systems have shown the remarkable effectiveness of deep learning techniques in this task, especially when aided by modern pretrained language models. Unfortunately, such systems are still not available as ready-to-use end-to-end packages, making it difficult for researchers to take advantage of their performance. The only alternative for a user interested in applying WSD to downstream tasks is to rely on currently available end-to-end WSD systems, which, however, still rely on graph-based heuristics or non-neural machine learning algorithms. In this paper, we fill this gap and propose AMuSE-WSD, the first end-to-end system to offer high-quality sense information in 40 languages through a state-of-the-art neural model for WSD. We hope that AMuSE-WSD will provide a stepping stone for the integration of meaning into real-world applications and encourage further studies in lexical semantics. AMuSE-WSD is available online at http://nlp.uniroma1.it/amuse-

wsd.

SummerTime: Text Summarization Toolkit for Non-experts
Ansong Ni et al.

Recent advances in summarization provide models that can generate summaries of higher quality. Such models now exist for a number of summarization tasks, including query-based summarization, dialogue summarization, and multi-document summarization. While such models and tasks are rapidly growing in the research field, it has also become challenging for non-experts to keep track of them. To make summarization methods more accessible to a wider audience, we develop SummerTime by rethinking the summarization task from the perspective of an NLP non-expert. SummerTime is a complete toolkit for text summarization, including various models, datasets, and evaluation metrics, for a full spectrum of summarization-related tasks. SummerTime integrates with libraries designed for NLP researchers, and enables users with easy-to-use APIs. With SummerTime, users can locate pipeline solutions and search for the best model with their own data, and visualize the differences, all with a few lines of code. We also provide explanations for models and evaluation metrics to help users understand the model behaviors and select models that best suit their needs. Our library, along with a notebook demo, is available at https://github.com/Yale-LILY/SummerTime.

Chandler: An Explainable Sarcastic Response Generator
Silviu Oprea, Steven Wilson, and Walid Magdy

We introduce Chandler, a system that generates sarcastic responses to a given utterance. Previous sarcasm generators assume the intended meaning that sarcasm conceals is the opposite of the literal meaning. We argue that this traditional theory of sarcasm provides a grounding that is neither necessary, nor sufficient, for sarcasm to occur. Instead, we ground our generation process on a formal theory that specifies conditions that unambiguously differentiate sarcasm from non-sarcasm. Furthermore, Chandler not only generates sarcastic responses, but also explanations for why each response is sarcastic. This provides accountability, crucial for avoiding miscommunication between humans and conversational agents, particularly considering that sarcastic communication can be offensive. In human evaluation, Chandler achieves comparable or higher sarcasm scores, compared to state-of-the-art generators, while generating more diverse responses, that are more specific and more coherent to the input.

DRIFT: A Toolkit for Diachronic Analysis of Scientific Literature
Abheesht Sharma et al.

In this work, we present to the NLP community, and to the wider research community as a whole, an application for the diachronic analysis of research corpora. We open source an easy-to-use tool coined DRIFT, which allows researchers to track research trends and development over the years. The analysis methods are collated from well-cited research works, with a few of our own methods added for good measure. Succinctly put, some of the analysis methods are: keyword extraction, word clouds, predicting declining/stagnant/growing trends using Productivity, tracking bi-grams using Acceleration plots, finding the Semantic Drift of words, tracking trends using similarity, etc. To demonstrate the utility and efficacy of our tool, we perform a case study on the cs.CL corpus of the arXiv repository and draw inferences from the analysis methods. The toolkit and the associated code are available here: https://github.com/rajaswa/DRIFT.

FAST: Fast Annotation tool for SmarT devices
Shunyo Kawamoto et al.

Working with a wide range of annotators with the same attributes is crucial, as in real-world applications. Although such application cases often use crowd-sourcing mechanisms to gather a variety of annotators, most real-world users use mobile devices. In this paper, we propose "FAST," an annotation tool for application tasks that focuses on the user experience of mobile devices, which has not yet been focused on thus far. We designed FAST as a web application for use on any device with a flexible interface that can be customized to fit various tasks. In our experiments, we conducted crowd-sourced annotation for a sentiment analysis task with several annotators and evaluated annotation metrics such as speed, quality, and ease of use from the tool's logs and user surveys. Based on the results of our experiments, we conclude that our system can annotate faster than existing methods while maintaining the annotation quality.

7 Workshops

Wednesday–Thursday

Wednesday

Thursday

Workshop 11: Sixth Conference on Machine Translation

Organizers: *Barry Haddow et al.*

Wednesday, November 10, 2021

9:00–9:10 **Opening Remarks**

9:10–10:15 **Session 1: Shared Task Overview Papers I**

9:10–10:15 **wmt-161**
wmt-161

9:00–9:25 **News Translation Task**

9:25–9:35 **Similar Languages Translation Task**

9:35–9:45 **Automatic Postediting Task**

9:45–9:55 **Triangular Translation Task**

9:55–10:05 **Indo-European Multilingual Translation Task**

10:15–10:30 **Coffee Break**

10:30–12:00 **News Translation Task**

10:30–12:00 **Similar Languages Translation Task**

10:30–12:00 **Automatic Post-Editing Task**

10:30–12:00 **Triangular Translation Task**

10:30–12:00 **Indo-European Multilingual Translation Task**

10:30–12:00 **Large-Scale Multilingual Translation Task**

12:00–1:00 **Lunch Break**

1:00–2:15 **Session 3: Panel Discussion on Evaluation with Nitika Mathur (Univ. Melbourne), Benjamin Marie (NICT), Ricardo Rei (Unbabel), Tom Kocmi (Microsoft) and moderated by Markus Freitag (Google)**

2:15–2:45 **Mini Break**

2:45–4:15 **Session 4: Research Papers on Evaluation**

4:15-4:45 **Coffee Break**

4:45–6:15 **Session 5: Research Papers on Data**

Thursday, November 11, 2021

9:00–10:15 **Session 6: Shared Task Overview Papers I**

10:15–10:30 **Coffee Break**

10:30–12:00 **Efficient Translation Task**

10:30–12:00 **Terminology Translation Task**

10:30–12:00 **Biomedical Translation Task**

10:30–12:00 **Quality Estimation Task**

10:30–12:00 **Unsupervised and Very Low Resource Translation Task**

10:30–12:00 **Metrics Task**

10:30–12:00 **Test Suites**

12:00–1:00 **Lunch Break**

1:00–2:15 **Session 8: Research Papers on Training and Modelling**

2:15–2:45 **Mini Break**

2:45–4:15 **Session 9: Machine Translation Papers from the Findings of the EMNLP**

4:15-4:45 **Coffee Break**

4:45–6:15 **Session 10: Machine Translation Papers from the Findings of the EMNLP**

Workshop 22: The 25th Conference on Computational Natural Language Learning (CoNLL)

Organizers: *Omri Abend et al.*

Wednesday, November 10, 2021

10:00–10:10 **Welcome**

10:10–11:30 **Oral session 1: Interaction, dialogue, and grounded language learning**

10:10–10:30 "It's our fault!": Insights Into Users' Understanding and Interaction With an Explanatory Collaborative Dialog System
Katharina Weitz et al.

10:30–10:50 Dependency Induction Through the Lens of Visual Perception
Ruisi Su et al.

10:50–11:10 VQA-MHUG: A Gaze Dataset to Study Multimodal Neural Attention in Visual Question Answering
Ekta Sood et al.

11:10–11:30 "It seemed like an annoying woman": On the Perception and Ethical Considerations of Affective Language in Text-Based Conversational Agents
Lindsey Vanderlyn et al.

11:30–12:00 **Break**

12:00–1:10 **Keynote I: Linking learning to language typology (Jennifer Culbertson)**

1:10–2:10 **Lunch break**

2:10–3:50 **Oral session 2: Theoretical analysis, probing, and interpretation of language models**

2:10–2:30 On Language Models for Creoles
Heather Lent et al.

2:30–2:50 Do pretrained transformers infer telicity like humans?
Yiyun Zhao et al.

2:50–3:10 The Low-Dimensional Linear Geometry of Contextualized Word Representations
Evan Hernandez and Jacob Andreas

3:10–3:30 Generalising to German Plural Noun Classes, from the Perspective of a Recurrent Neural Network
Verna Dankers et al.

3:30–3:50 Can Language Models Encode Perceptual Structure Without Grounding? A Case Study in Color
Mostafa Abdou et al.

3:50–4:20 **Break**

4:20–6:00 **Poster session 1**

4:20–6:00 Empathetic Dialog Generation with Fine-Grained Intents
Yubo Xie and Pearl Pu

4:20–6:00 Enriching Language Models with Visually-grounded Word Vectors and the Lancaster Sensorimotor Norms
Casey Kennington

4:20–6:00 Learning Zero-Shot Multifaceted Visually Grounded Word Embeddings via Multi-Task Training
Hassan Shahmohammadi, Hendrik P. A. Lensch, and R. Harald Baayen

4:20–6:00 Does language help generalization in vision models?
Benjamin Devillers et al.

4:20–6:00 Understanding Guided Image Captioning Performance across Domains
Edwin G. Ng et al.

4:20–6:00 Counterfactual Interventions Reveal the Causal Effect of Relative Clause Representations on Agreement Prediction
Shauli Ravfogel et al.

4:20–6:00 Who's on First?: Probing the Learning and Representation Capabilities of Language Models on Deterministic Closed Domains
David Demeter and Doug Downey

4:20–6:00 Data Augmentation of Incorporating Real Error Patterns and Linguistic Knowledge for Grammatical Error Correction
Xia Li and Junyi He

4:20–6:00 Agree to Disagree: Analysis of Inter-Annotator Disagreements in Human Evaluation of Machine Translation Output
Maja Popović

4:20–6:00 A Multilingual Benchmark for Probing Negation-Awareness with Minimal Pairs
Mareike Hartmann et al.

4:20–6:00 Explainable Natural Language to Bash Translation using Abstract Syntax Tree
Shikhar Bharadwaj and Shirish Shevade

4:20–6:00 Learned Construction Grammars Converge Across Registers Given Increased Exposure
Jonathan Dunn and Harish Tayyar Madabushi

4:20–6:00 Tokenization Repair in the Presence of Spelling Errors
Hannah Bast, Matthias Hertel, and Mostafa M. Mohamed

4:20–6:00 A Coarse-to-Fine Labeling Framework for Joint Word Segmentation, POS Tagging, and Constituent Parsing
Yang Hou et al.

4:20–6:00 Understanding the Extent to which Content Quality Metrics Measure the Information Quality of Summaries
Daniel Deutsch and Dan Roth

Thursday, November 11, 2021

10:00–11:40 **Oral session 3: Lexical, compositional, and discourse semantics; Pragmatics**

10:00–10:20 Summary-Source Proposition-level Alignment: Task, Datasets and Supervised Baseline
Ori Ernst et al.

10:20–10:40 Exploring Metaphoric Paraphrase Generation
Kevin Stowe, Nils Beck, and Iryna Gurevych

10:40–11:00 Imposing Relation Structure in Language-Model Embeddings Using Contrastive Learning
Christos Theodoropoulos et al.

11:00–11:20 NOPE: A Corpus of Naturally-Occurring Presuppositions in English
Alicia Parrish et al.

11:20–11:40 Pragmatic competence of pre-trained language models through the lens of discourse connectives
Lalchand Pandia, Yan Cong, and Allyson Ettinger

11:40–1:20 **Poster session 2**

11:40–1:20 Predicting Text Readability from Scrolling Interactions
Sian Gooding et al.

11:40–1:20 Modeling the Interaction Between Perception-Based and Production-Based Learning in Children's Early Acquisition of Semantic Knowledge
Mitja Nikolaus and Abdellah Fourtassi

11:40–1:20 Scaffolded input promotes atomic organization in the recurrent neural network language model
Philip A. Huebner and Jon A. Willits

11:40–1:20 Grammatical Profiling for Semantic Change Detection
Andrey Kutuzov, Lidia Pivovarova, and Mario Giulianelli

11:40–1:20 Deconstructing syntactic generalizations with minimalist grammars
Marina Ermolaeva

11:40–1:20 Relation-aware Bidirectional Path Reasoning for Commonsense Question Answering
Junxing Wang et al.

11:40–1:20 Does referent predictability affect the choice of referential form? A computational approach using masked coreference resolution
Laura Aina et al.

11:40–1:20 Polar Embedding
Ran Iwamoto, Ryosuke Kohita, and Akifumi Wachi

11:40–1:20 Commonsense Knowledge in Word Associations and ConceptNet
Chunhua Liu, Trevor Cohn, and Lea Frermann

11:40–1:20 Cross-document Event Identity via Dense Annotation
Adithya Pratapa et al.

11:40–1:20 Tackling Zero Pronoun Resolution and Non-Zero Coreference Resolution Jointly
Shisong Chen et al.

11:40–1:20 Negation-Instance Based Evaluation of End-to-End Negation Resolution
Elizaveta Sineva et al.

11:40–1:20 Controlling Prosody in End-to-End TTS: A Case Study on Contrastive Focus Generation
Siddique Latif et al.

11:40–1:20 A Large-scale Comprehensive Abusiveness Detection Dataset with Multifaceted Labels from Reddit
Hoyun Song et al.

11:40–1:20 MirrorWiC: On Eliciting Word-in-Context Representations from Pretrained Language Models
Qianchu Liu et al.

11:40–1:20 A Data Bootstrapping Recipe for Low-Resource Multilingual Relation Classification
Arijit Nag et al.

11:40–1:20 FAST: A carefully sampled and cognitively motivated dataset for distributional semantic evaluation
Stefan Evert and Gabriella Lapesa

11:40–1:20 Automatic Error Type Annotation for Arabic
Riadh Belkebir and Nizar Habash

1:20–2:20 **Lunch break**

2:20–3:30 **Keynote II: What are we learning from language? (Gary Lupyan)**

3:30–3:50 **Break**

3:50–4:50 **Oral session 4: Language evolution, acquisition and linguistic theories**

3:50–4:10 The Emergence of the Shape Bias Results from Communicative Efficiency
Eva Portelance et al.

4:10–4:30 BabyBERTa: Learning More Grammar With Small-Scale Child-Directed Language
Philip A. Huebner et al.

4:30–4:50 Analysing Human Strategies of Information Transmission as a Function of Discourse Context
Mario Giulianelli and Raquel Fernández

4:50–5:10 **Break**

5:10–5:50 **Oral session 5: Speech and phonology**

5:10–5:30 Predicting non-native speech perception using the Perceptual Assimilation Model and state-of-the-art acoustic models
Juliette Millet, Ioana Chitoran, and Ewan Dunbar

5:30–5:50 The Influence of Regional Pronunciation Variation on Children's Spelling and the Potential Benefits of Accent Adapted Spellcheckers
Emma O'Neill et al.

5:50–6:20 **Best Paper Award and Closing Words**

Workshop 9: 8th International Workshop on Argument Mining

Organizers: *Khalid Al Khatib, Yufang Hou, and Manfred Stede*

Wednesday, November 10, 2021

9:00–9:30 **Opening remarks**

9:30–10:30 **Invited Talk 1**

10:30–11:00 **Coffee Break**

11:00–12:00 **Session 1**

- Multi-task and Multi-corpora Training Strategies to Enhance Argumentative Sentence Linking Performance
  Jan Wira Gotama Putra, Simone Teufel, and Takenobu Tokunaga
- Argument Mining on Twitter: A Case Study on the Planned Parenthood Debate
  Muhammad Mahad Afzal Bhatti, Ahsan Suheer Ahmad, and Joonsuk Park
- Explainable Unsupervised Argument Similarity Rating with Abstract Meaning Representation and Conclusion Generation
  Juri Opitz et al.

12:00–1:00 **Lunch Break**

1:00–2:30 **Session 2**

- Multilingual Counter Narrative Type Classification
  Yi-Ling Chung, Marco Guerini, and Rodrigo Agerri
- Predicting Moderation of Deliberative Arguments: Is Argument Quality the Key?
  Neele Falk et al.
- Self-trained Pretrained Language Models for Evidence Detection
  Diane Litman
- Multi-task Learning in Argument Mining for Persuasive Online Discussions
  Nhat Tran and Diane Litman

2:30–2:45 **Coffee Break**

2:45–4:15 **Panel Talks and Discussion**

4:15–4:45 **Coffee Break**

4:45–5:45 **Session 3**

- Image Retrieval for Arguments Using Stance-Aware Query Expansion
  Johannes Kiesel et al.
- Is Stance Detection Topic-Independent and Cross-topic Generalizable? - A Reproduction Study
  Myrthe Reuver et al.
- Exploring Methodologies for Collecting High-Quality Implicit Reasoning in Arguments
  Keshav Singh et al.

Thursday, November 11, 2021

9:00–10:00 **Invited Talk 2**

10:00–10:40 **Session 4**

- Assessing the Sufficiency of Arguments through Conclusion Generation
  Timon Gurcke, Milad Alshomary, and Henning Wachsmuth
- M-Arg: Multimodal Argument Mining Dataset for Political Debates with Audio and Transcripts
  Rafael Mestre et al.

10:40–11:00 **Coffee Break**

11:00–12:00 **Session 5**

- Citizen Involvement in Urban Planning - How Can Municipalities Be Supported in Evaluating Public Participation Processes for Mobility Transitions?
  Julia Romberg and Stefan Conrad
- Argumentation Mining in Scientific Literature for Sustainable Development
  Aris Fergadis et al.
- Bayesian Argumentation-Scheme Networks: A Probabilistic Model of Argument Validity Facilitated by Argumentation Schemes
  Takahiro Kondo et al.

12:00–1:00 **Lunch Break**

1:00–2:30 **Shared Task Presentation**

- Overview of the 2021 Key Point Analysis Shared Task
  Roni Friedman et al.
- Matching The Statements: A Simple and Accurate Model for Key Point Analysis
  Hoang Phan et al.
- Modern Talking in Key Point Analysis: Key Point Matching using Pretrained Encoders
  Jan Heinrich Reimer et al.
- Key Point Analysis via Contrastive Learning and Extractive Argument Summarization
  Milad Alshomary et al.
- Key Point Matching with Transformers
  Emanuele Cosenza
- Team Enigma at ArgMining-EMNLP 2021: Leveraging Pre-trained Language Models for Key Point Matching
  Manav Kapadnis et al.
- Key Point Analysis with a siamese transformer
  Jan Bittner and Johannes Huck

2:30–2:45 **Break**

2:45–3:00 **Concluding Remarks**

Workshop 16: 2nd Workshop on Computational Approaches to Discourse (CODI)

Organizers: *Chloé Braud et al.*

Wednesday, November 10, 2021

9:00–12:00 **CODI-CRAC Shared Task**

9:05–9:30 **Welcome**

9:30–9:45 Neural Anaphora Resolution in Dialogues (Hideo Kobayashi, Shengjie Li and Vincent Ng)

9:30–9:45 Neural Anaphora Resolution in Dialogues
Hideo Kobayashi, Shengjie Li and Vincent Ng

9:45–10:00 **Anaphora Resolution in Dialogue: Description of the DFKI-TalkingRobots System for the CODI-CRAC 2021 Shared-Task (Tatiana Anikina, Cennet Oguz, Natalia Skachkova, Siyu Tao, Sharmila Upadhyaya and Ivana Kruijff-Korbayova)**

10:00–10:30 **Coffee Break**

10:30–10:45 The Pipeline Model for Resolution of Anaphoric Reference and Resolution of Entity Reference (Hongjin Kim, Damrin Kim and Harksoo Kim)

10:45–11:00 An End-to-End Approach for Full Bridging Resolution (Joseph Renner, Priyansh Trivedi, Gaurav Maheshwari, Rémi Gilleron and Pascal Denis)

11:00–11:15 Adapted End-to-End Coreference Resolution System for Anaphoric Identities in Dialogues (Liyan Xu and Jinho D. D. Choi)

11:15–11:30 Anaphora Resolution in Dialogue: Cross-Team Analysis of the DFKI-TalkingRobots Team Submissions for the CODI-CRAC 2021 Shared-Task (Natalia Skachkova, Cennet Oguz, Tatiana Anikina, Siyu Tao, Sharmila Upadhyaya and Ivana Kruijff-Korbayova)

11:30–11:45 The CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis Resolution in Dialogue: A Cross-Team Analysis (Shengjie Li, Hideo Kobayashi, and Vincent Ng)

11:45–12:00 **Plenary Session**

11:45–12:00 Visioning Discussion and Next Steps

12:00–1:30 **Lunch Break**

1:30–2:30 **Plenary Session**

1:30–2:30 **Invited Talk (Jackie Chi Kit Cheung)**

2:30–2:45 **Mini Break**

2:45–3:45 **Pragmatics and Applications**

2:45–3:00 "I'll be there for you": The One with Understanding Indirect Answers
Cathrine Damgaard et al.

3:00–3:10 Improving Text Generation via Neural Discourse Planning
Alexander Chernyavskiy and Dmitry Ilvovsky

3:10–3:25 Developing Conversational Data and Detection of Conversational Humor in Telugu
Vaishnavi Pamulapati and Radhika Mamidi

3:25–3:35 Investigating non lexical markers of the language of schizophrenia in spontaneous conversations
Chuyuan Li et al.

3:35–3:45 Discourse-Driven Integrated Dialogue Development Environment for Open-Domain Dialogue Systems
Denis Kuznetsov et al.

3:50–4:15 **Anaphora and Coreference**

3:50–4:00 Coreference Chains Categorization by Sequence Clustering
Silvia Federzoni, Lydia-Mai Ho-Dac, and Cécile Fabre

4:00–4:15 Resolving Implicit References in Instructional Texts
Talita Anthonio and Michael Roth

4:15–4:45 **Coffee Break**

4:45–6:15 **Discourse Relations**

4:45–5:00 A practical perspective on connective generation
Frances Yung, Merel Scholman, and Vera Demberg

5:00–5:15 Semi-automatic discourse annotation in a low-resource language: Developing a connective lexicon for Nigerian Pidgin
Marian Marchal, Merel Scholman, and Vera Demberg

5:15–5:30 Comparison of methods for explicit discourse connective identification across various domains
Merel Scholman et al.

5:30–5:40 A Novel Corpus of Discourse Structure in Humans and Computers
Babak Hemmatian et al.

5:40–5:55 Revisiting Shallow Discourse Parsing in the PDTB-3: Handling Intra-sentential Implicits
Zheng Zhao and Bonnie Webber

5:55–6:05 Improving Multi-Party Dialogue Discourse Parsing via Domain Integration
Zhengyuan Liu and Nancy Chen

6:05–6:15 discopy: A Neural System for Shallow Discourse Parsing
René Knaebel

Thursday, November 11, 2021

9:00–10:00 **Plenary Session**

9:00–10:00 **Invited Talk: Inter annotator agreement in discourse annotation – the role of domain knowledge and individual differences (Vera Demberg)**

10:25–11:15 **Discourse and Multilinguality**

10:25–10:35 Tracing variation in discourse connectives in translation and interpreting through neural semantic spaces
Ekaterina Lapshinova-Koltunski, Heike Przybyl, and Yuri Bizzoni

10:35–10:50 Capturing document context inside sentence-level neural machine translation models with self-training
Elman Mansimov, Gábor Melis, and Lei Yu

10:50–11:05 DMRST: A Joint Framework for Document-Level Multilingual RST Discourse Segmentation and Parsing
Zhengyuan Liu, Ke Shi, and Nancy Chen

11:05–11:15 Visualizing CrossLingual Discourse Relations in Multilingual TED Corpora
Zae Myung Kim et al.

11:15–12:00 **Discussion and Closing of Main CODI Workshop**

12:00–1:30 **Lunch Break**

1:30–3:30 **DISRPT 2021 Shared Task Session 1**

1:50–2:20 A Transformer Based Approach towards Identification of Discourse Unit Segments and Connectives (Sahil Bakshi and Dipti Misra Sharma)

2:20–2:50 Multi-lingual Discourse Segmentation and Connective Identification: MELODI at DISRPT2021 (Morteza Ezzabady, Philippe Muller, and Chloé Braud)

2:50–3:20 Delexicalised Multilingual Discourse Segmentation for DISRPT 2021 and Tense, Mood, Voice and Modality Tagging for 11 Languages (Tillmann D´onicke)

3:30–4:00 **Coffee Break**

4:00–5:30 **DISRPT 2021 Shared Task Session 2**

4:00–4:30 A Unified Approach to Discourse Relation Classification in nine Languages (Hanna Varachkina and Franziska Pannach)

4:30–5:00 DisCoDisCo at the DISRPT2021 Shared Task: A System for Discourse Segmentation, Classification, and Connective Detection (Luke Gessler, Shabnam Behzad, Yang Janet Liu, Siyao Peng, Yilun Zhu, Amir Zeldes)

5:00–5:15 **Closing Remarks**

Workshop 3: 3rd Workshop on NLP for Conversational AI (NLP4ConvAI)

Organizers: *Alexandros Papangelis et al.*

- Taking Things Personally: Third Person to First Person Rephrasing
  Marcel Granero Moya and Panagiotis Agis Oikonomou Filandras
- Few-Shot Intent Classification by Gauging Entailment Relationship Between Utterance and Semantic Label
  Jin Qu et al.
- Personalized Extractive Summarization Using an Ising Machine Towards Real-time Generation of Efficient and Coherent Dialogue Scenarios
  Hiroaki Takatsu et al.
- Multilingual Paraphrase Generation For Bootstrapping New Features in Task-Oriented Dialog Systems
  Subhadarshi Panda et al.
- Overcoming Conflicting Data when Updating a Neural Semantic Parser
  David Gaddy et al.
- Not So Fast, Classifier — Accuracy and Entropy Reduction in Incremental Intent Classification
  Lianna Hrycyk, Alessandra Zarcone, and Luzian Hahn
- On the Robustness of Intent Classification and Slot Labeling in Goal-oriented Dialog Systems to Real-world Noise
  Sailik Sengupta, Jason Krone, and Saab Mansour
- Amendable Generation for Dialogue State Tracking
  Xin Tian et al.
- What Went Wrong? Explaining Overall Dialogue Quality through Utterance-Level Impacts
  James D. Finch, Sarah E. Finch, and Jinho D. Choi
- XPersona: Evaluating Multilingual Personalized Chatbot
  Zhaojiang Lin et al.
- Collaborative Data Relabeling for Robust and Diverse Voice Apps Recommendation in Intelligent Personal Assistants
  Qian Hu et al.
- Semi-supervised Intent Discovery with Contrastive Learning
  Xiang Shen et al.
- CS-BERT: a pretrained model for customer service dialogues
  Peiyao Wang, Joyce Fang, and Julia Reinspach
- PLATO-KAG: Unsupervised Knowledge-Grounded Conversation via Joint Modeling
  Xinxian Huang et al.
- Improving Dialogue State Tracking by Joint Slot Modeling
  Ting-Rui Chiang and Yi-Ting Yeh
- Learning to Learn End-to-End Goal-Oriented Dialog From Related Dialog Tasks
  Janarthanan Rajendran, Jonathan K. Kummerfeld, and Satinder Baveja
- Personalized Search-based Query Rewrite System for Conversational AI
  Eunah Cho et al.
- Dialogue Response Generation via Contrastive Latent Representation Learning
  Shuyang Dai et al.
- AuGPT: Auxiliary Tasks and Data Augmentation for End-To-End Dialogue with Pre-Trained Language Models
  Jonáš Kulhánek et al.
- Investigating Pretrained Language Models for Graph-to-Text Generation
  Leonardo F. R. Ribeiro et al.
- Style Control for Schema-Guided Natural Language Generation
  Alicia Tsai et al.
- Using Pause Information for More Accurate Entity Recognition
  Sahas Dendukuri et al.
- Think Before You Speak: Learning to Generate Implicit Knowledge for Response Generation by Self-Talk
  Pei Zhou et al.
- Teach Me What to Say and I Will Learn What to Pick: Unsupervised Knowledge Selection Through Response Generation with Pretrained Generative Models
  Ehsan Lotfi et al.

- Influence of user personality on dialogue task performance: A case study using a rule-based dialogue system
  Ao Guo et al.
- Towards Code-Mixed Hinglish Dialogue Generation
  Vibhav Agarwal, Pooja Rao, and Dinesh Babu Jayagopi
- Towards Zero and Few-shot Knowledge-seeking Turn Detection in Task-orientated Dialogue Systems
  Di Jin et al.

Workshop 4: Novel Ideas in Learning-to-Learn through Interaction

Organizers: *Prasanna Parthasarathi et al.*

12:00–12:15 **Opening Remarks (Prasanna Parthasarathi)**
12:15–1:00 **Invited Talk 1 (Prof. Tom Mitchell)**
1:00–2:00 **Poster Session 1**
2:00–2:45 **Invited Talk 2 (Prof. Yejin Choi)**
2:45–3:15 **Contributed Talks**
4:15–5:00 **Invited Talk 3 (Dr. Felix Hill)**
5:00–5:45 **Invited Talk 4 (Dr. Douwe Kiela)**
5:45–6:45 **Poster Session 2**
6:45–7:30 **Invited Talk 5 (Dr. Natasha Jaques)**
7:30–8:30 **Contributed Talks**
8:30–9:15 **Panel Discussion**
9:15–9:30 **Concluding Remarks**

Workshop 5: 2nd Workshop on Evaluation and Comparison of NLP Systems (Eval4NLP)

Organizers: *Steffen Eger et al.*

9:00–9:10 **Opening Remarks**

9:15–9:55 **Keynote Talk 1**

10:00–10:40 **Paper Presentation Session 1**

- Statistically Significant Detection of Semantic Shifts using Contextual Word Embeddings
  Yang Liu, Alan Medlar, and Dorota Glowacka
- Developing a Benchmark for Reducing Data Bias in Authorship Attribution
  Benjamin Murauer and Günther Specht
- ESTIME: Estimation of Summary-to-Text Inconsistency by Mismatched Embeddings
  Oleg Vasilyev and John Bohannon
- Testing Cross-Database Semantic Parsers With Canonical Utterances
  Heather Lent et al.

10:45–11:25 **Keynote Talk 2**

11:30–12:10 **Paper Presentation Session 2**

- Writing Style Author Embedding Evaluation
  Enzo Terreau, Antoine Gourru, and Julien Velcin
- HinGE: A Dataset for Generation and Evaluation of Code-Mixed Hinglish Text
  Vivek Srivastava and Mayank Singh
- Error-Sensitive Evaluation for Ordinal Target Variables
  David Chen et al.
- MIPE: A Metric Independent Pipeline for Effective Code-Mixed NLG Evaluation
  Ayush Garg et al.

12:15–12:55 **Keynote Talk 3**

1:00–2:00 **Lunch Break**

2:00–2:40 **Keynote Talk 4**

2:45–3:25 **Paper Presentation Session 3**

- SeqScore: Addressing Barriers to Reproducible Named Entity Recognition Evaluation
  Chester Palen-Michel, Nolan Holley, and Constantine Lignos
- How Emotionally Stable is ALBERT? Testing Robustness with Stochastic Weight Averaging on a Sentiment Analysis Task
  Urja Khurana, Eric Nalisnick, and Antske Fokkens
- StoryDB: Broad Multi-language Narrative Dataset
  Alexey Tikhonov, Igor Samenko, and Ivan Yamshchikov
- What is SemEval evaluating? A Systematic Analysis of Evaluation Campaigns in NLP
  Oskar Wysocki et al.
- Differential Evaluation: a Qualitative Analysis of Natural Language Processing System Behavior Based Upon Data Resistance to Processing
  Lucie Gianola et al.

3:30–4:10 **Keynote Talk 5**

4:15–4:55 **Paper Presentation Session 4**

- Trainable Ranking Models to Evaluate the Semantic Accuracy of Data-to-Text Neural Generator
  Nicolas Garneau and Luc Lamontagne
- Referenceless Parsing-Based Evaluation of AMR-to-English Generation
  Emma Manning and Nathan Schneider
- Evaluation of Unsupervised Automatic Readability Assessors Using Rank Correlations
  Yo Ehara
- Validating Label Consistency in NER Data Annotation
  Qingkai Zeng et al.

5:00–5:45 **Award Announcment Shared Task Winners Presentation**

- The Eval4NLP Shared Task on Explainable Quality Estimation: Overview and Results
  Marina Fomicheva et al.

- IST-Unbabel 2021 Submission for the Explainable Quality Estimation Shared Task
  Marcos Treviso et al.
- The UMD Submission to the Explainable MT Quality Estimation Shared Task: Combining Explanation Models with Sequence Labeling
  Tasnim Kabir and Marine Carpuat
- Reference-Free Word- and Sentence-Level Translation Evaluation with Token-Matching Metrics
  Christoph Wolfgang Leiter
- Explaining Errors in Machine Translation with Absolute Gradient Ensembles
  Melda Eksi et al.
- Explainable Quality Estimation: CUNI Eval4NLP Submission
  Peter Polák, Muskaan Singh, and Ondřej Bojar
- Error Identification for Machine Translation with Metric Embedding and Attention
  Raphael Rubino, Atsushi Fujita, and Benjamin Marie

5:50–6:00 **Concluding Remarks**

Workshop 6: CI+NLP: First Workshop on Causal Inference in NLP

Organizers: *Jacob Eisenstein et al.*

9:00–9:10 **Opening Remarks**

9:10–10:00 **Keynote 1**

10:00–10:30 **Coffee Break**

10:30–12:00 **Short Talks Session 1**

10:50–11:00 Causal Augmentation for Causal Sentence Classification
Fiona Anting Tan et al.

11:00–11:10 Text as Causal Mediators: Research Design for Causal Estimates of Differential Treatment of Social Groups via Language Aspects
Katherine Keith, Douglas Rice, and Brendan O'Connor

11:10–11:20 Enhancing Model Robustness and Fairness with Causality: A Regularization Approach
Zhao Wang, Kai Shu, and Aron Culotta

11:20–11:30 What Makes a Scientific Paper be Accepted for Publication?
Panagiotis Fytas, Georgios Rizos, and Lucia Specia

12:00–1:00 **Lunch Break**

1:00–2:30 **Short Talks Session 2**

1:10–1:20 Sensitivity Analysis for Causal Mediation through Text: an Application to Political Polarization
Graham Tierney and Alexander Volfovsky

1:20–1:30 A Survey of Online Hate Speech through the Causal Lens
Antigoni Founta and Lucia Specia

1:30–1:40 Identifying Causal Influences on Publication Trends and Behavior: A Case Study of the Computational Linguistics Community
Maria Glenski and Svitlana Volkova

1:40–1:50 It's quality and quantity: the effect of the amount of comments on online suicidal posts
Daniel Low et al.

2:30–2:45 **Mini Break**

2:45–3:35 **Invited Talk 2**

3:35–4:25 **Invited Talk 3**

4:25–4:45 **Coffee Break**

4:45–5:30 **Panel Discussion**

5:30–6:30 **Poster Session**

Workshop 10: The Second Workshop on Insights from Negative Results in NLP

Organizers: *João Sedoc et al.*

- Corrected CBOW Performs as well as Skip-gram
  Ozan İrsoy, Adrian Benton, and Karl Stratos
- Does Commonsense help in detecting Sarcasm?
  Somnath Basu Roy Chowdhury and Snigdha Chaturvedi
- BERT Cannot Align Characters
  Antonis Maronikolakis, Philipp Dufter, and Hinrich Schütze
- Two Heads are Better than One? Verification of Ensemble Effect in Neural Machine Translation
  Sungjin Park et al.
- Finetuning Pretrained Transformers into Variational Autoencoders
  Seongmin Park and Jihwa Lee
- Are BERTs Sensitive to Native Interference in L2 Production?
  Zixin Tang, Prasenjit Mitra, and David Reitter
- Zero-Shot Cross-Lingual Transfer is a Hard Baseline to Beat in German Fine-Grained Entity Typing
  Sabine Weber and Mark Steedman
- Comparing Euclidean and Hyperbolic Embeddings on the WordNet Nouns Hypernymy Graph
  Sameer Bansal and Adrian Benton
- When does Further Pre-training MLM Help? An Empirical Study on Task-Oriented Dialog Pre-training
  Qi Zhu et al.
- Recurrent Attention for the Transformer
  Jan Rosendahl et al.
- On the Difficulty of Segmenting Words with Attention
  Ramon Sanabria, Hao Tang, and Sharon Goldwater
- The Highs and Lows of Simple Lexical Domain Adaptation Approaches for Neural Machine Translation
  Nikolay Bogoychev and Pinzhen Chen
- Backtranslation in Neural Morphological Inflection
  Ling Liu and Mans Hulden
- Learning Data Augmentation Schedules for Natural Language Processing
  Daphné Chopard, Matthias S. Treder, and Irena Spasić
- An Investigation into the Contribution of Locally Aggregated Descriptors to Figurative Language Identification
  Sina Mahdipour Saravani, Ritwik Banerjee, and Indrakshi Ray
- Blindness to Modality Helps Entailment Graph Mining
  Liane Guillou et al.
- Investigating the Effect of Natural Language Explanations on Out-of-Distribution Generalization in Few-shot NLI
  Yangqiaoyu Zhou and Chenhao Tan
- Generalization in NLI: Ways (Not) To Go Beyond Simple Heuristics
  Prajjwal Bhargava, Aleksandr Drozd, and Anna Rogers
- Challenging the Semi-Supervised VAE Framework for Text Classification
  Ghazi Felhi, Joseph Le Roux, and Djamé Seddah
- Active Learning for Argument Strength Estimation
  Nataliia Kees et al.

Workshop 12: The 3rd Workshop on Machine Reading for Question Answering

Organizers: *Adam Fisch et al.*

Wednesday, November 10, 2021

9:00–9:15 **Opening Remarks**

9:15–11:30 **Multilingual QA Invited Talk Session**

9:15–9:45 **Invited Talk 1 - Reut Tsarfaty**

9:45–10:15 **Invited Talk 2 - Jon Clark**

10:15–10:45 **Invited Talk 3 - Yiming Cui**

10:45–11:30 **Panel Discussion on Multilingual QA**

11:30–12:30 **Lunch break**

12:30–1:10 **Best Paper Talk Session**

12:30–12:44 MFAQ: a Multilingual FAQ Dataset
Maxime De Bruyn et al.

12:44–12:57 Rethinking the Objectives of Extractive Question Answering
Martin Fajcik, Josef Jon, and Pavel Smrz

12:57–1:10 What Would it Take to get Biomedical QA Systems into Practice?
Gregory Kell et al.

1:10–2:10 **Poster Session (archival track)**

1:10–2:10 GermanQuAD and GermanDPR: Improving Non-English Question Answering and Passage Retrieval
Timo Möller, Julian Risch, and Malte Pietsch

1:10–2:10 Zero-Shot Clinical Questionnaire Filling From Human-Machine Interactions
Farnaz Ghassemi Toudeshki et al.

1:10–2:10 Can Question Generation Debias Question Answering Models? A Case Study on Question—Context Lexical Overlap
Kazutoshi Shinoda, Saku Sugawara, and Akiko Aizawa

1:10–2:10 What Can a Generative Language Model Answer About a Passage?
Douglas Summers-Stay, Claire Bonial, and Clare Voss

1:10–2:10 Multi-modal Retrieval of Tables and Texts Using Tri-encoder Models
Bogdan Kostić, Julian Risch, and Timo Möller

1:10–2:10 Eliciting Bias in Question Answering Models through Ambiguity
Andrew Mao et al.

1:10–2:10 Bilingual Alignment Pre-Training for Zero-Shot Cross-Lingual Transfer
Ziqing Yang et al.

1:10–2:10 ParaShoot: A Hebrew Question Answering Dataset
Omri Keren and Omer Levy

1:10–2:10 Unsupervised Multiple Choices Question Answering: Start Learning from Basic Knowledge
Chi-Liang Liu and Hung-yi Lee

1:10–2:10 GANDALF: a General Character Name Description Dataset for Long Fiction
Fredrik Carlsson et al.

1:10–2:10 Investigating Post-pretraining Representation Alignment for Cross-Lingual Question Answering
Fahim Faisal and Antonios Anastasopoulos

1:10–2:10 Semantic Answer Similarity for Evaluating Question Answering Models
Julian Risch et al.

1:10–2:10 Simple and Efficient ways to Improve REALM
Vidhisha Balachandran et al.

1:10–2:10 **Poster Session (non-archival track)**

1:10–2:10 Synthetic Target Domain Supervision for Open Retrieval QA (Revanth Gangi Reddy, Bhavani Iyer, Md Arafat Sultan, Rong Zhang, Avirup Sil, Vittorio Castelli, Radu Florian, Salim Roukos)

1:10–2:10 Entity-based Knowledge Conflicts in Question Answering (Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris Dubois, Sameer Singh)

1:10–2:10 Mitigating False-Negative Contexts in Multi-Document Question Answering with Retrieval Marginalization (Ansong Ni, Matt Gardner, Pradeep Dasigi)

1:10–2:10 Generative Context Pair Selection for Multi-hop Question Answering (Dheeru Dua,Cicero Nogueira dos Santos,Patrick Ng,Ben Athiwaratkun,Bing Xiang,Matt Gardner,Sameer Singh)

1:10–2:10 Learning with Instance Bundles for Reading Comprehension (Dheeru Dua, Pradeep Dasigi,Sameer Singh,Matt Gardner)

1:10–2:10 Can NLI Models Verify QA Systems' Predictions? (Jifan Chen, Eunsol Choi, Greg Durrett)

1:10–2:10 Knowing More About Questions Can Help: Improving Calibration in Question Answering (Shujian Zhang, Chengyue Gong and Eunsol Choi)

1:10–2:10 RocketQA: An Optimized Training Approach to Dense Passage Retrieval for Open-Domain Question Answering (Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu and Haifeng Wang)

1:10–2:10 Weakly Supervised Pre-Training for Multi-Hop Retriever (Yeon Seonwoo, Sang-Woo Lee, Ji-Hoon Kim, Jung-Woo Ha and Alice Oh)

1:10–2:10 ReasonBert: Pre-trained to Reason with Distant Supervision (Xiang Deng, Yu Su, Alyssa Lees, You Wu, Cong Yu and Huan Sun)

1:10–2:10 Question Answering over Electronic Devices: A New Benchmark Dataset and a Multi-Task Learning based QA Framework (Abhilash Nandy, Soumya Sharma, Shubham Maddhashiya, Kapil Sachdeva, Pawan Goyal and NIloy Ganguly)

1:10–2:10 Do We Know What We Don't Know? Studying Unanswerable Questions beyond SQuAD 2.0 (Elior Sulem, Jamaal Hay and Dan Roth)

1:10–2:10 Relation-Guided Pre-Training for Open-Domain Question Answering (Ziniu Hu, Yizhou Sun and Kai-Wei Chang)

1:10–2:10 Beyond Reptile: Meta-Learned Dot-Product Maximization between Gradients for Improved Single-Task Regularization (Akhil Kedia, Sai Chetan Chinthakindi and Wonho Ryu)

1:10–2:10 SD-QA: Spoken Dialectal Question Answering for the Real World (Fahim Faisal, Sharlina Keshava, Md Mahfuz Ibn Alam and Antonios Anastasopoulos)

1:10–2:10 When Retriever-Reader Meets Scenario-Based Multiple-Choice Questions (ZiXian Huang, Ao Wu, Yulin Shen, Gong Cheng and Yuzhong Qu)

1:10–2:10 Winnowing Knowledge for Multi-choice Question Answering (Yeqiu Li, Bowei Zou, Zhifeng Li, Ai Ti Aw, Yu Hong and Qiaoming Zhu)

1:10–2:10 Extract, Integrate, Compete: Towards Verification Style Reading Comprehension (Chen Zhang, Yuxuan Lai, Yansong Feng and Dongyan Zhao)

1:10–2:10 Reference-based Weak Supervision for Answer Sentence Selection using Web Data (Vivek Krishnamurthy, Thuy Vu and Alessandro Moschitti)

1:10–2:10 NOAHQA: Numerical Reasoning with Interpretable Graph Question Answering Dataset (Qiyuan Zhang, Lei Wang, SICHENG YU, Shuohang Wang, Yang Wang, Jing Jiang and Ee-Peng Lim)

1:10–2:10 Improving Numerical Reasoning Skills in the Modular Approach for Complex Question Answering on Text (Xiao-Yu Guo, Yuan-Fang Li and Gholamreza Haffari)

1:10–2:10 R2-D2: A Modular Baseline for Open-Domain Question Answering (Martin Fajcik, Martin Docekal, Karel Ondrej and Pavel Smrz)

1:10–2:10 AutoEQA: Auto-Encoding Questions for Extractive Question Answering (Stalin Varanasi, Saadullah Amin and Guenter Neumann)

2:10–2:30 **Break**

2:30–4:45 **Interpretability in QA Invited Talk Session**

2:30–3:00 **Invited Talk 4 - Jonathan Berant**

3:00–3:30 **Invited Talk 5 - Marco Tulio Ribeiro**

3:30–4:00 **Invited Talk 6 - Hannaneh Hajishirzi**

4:00–4:45 **Panel Discussion on Interpretability in QA**

4:45–5:00 **Closing Remarks**

Workshop 15: The Natural Legal Language Processing Workshop 2021 (NLLP)

Organizers: *Nikolaos Aletras et al.*

Wednesday, November 10, 2021

8:00–8:10 **Workshop Opening**

8:10–9:00 **Invited Talk: John Armour**

9:00–9:15 A Corpus for Multilingual Analysis of Online Terms of Service
Kasper Drawzeski et al.

9:15–9:30 Named Entity Recognition in the Romanian Legal Domain
Vasile Pais et al.

9:45–10:00 Swiss-Judgment-Prediction: A Multilingual Legal Judgment Prediction Benchmark
Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer

10:45–11:00 Automated Extraction of Sentencing Decisions from Court Cases in the Hebrew Language
Mohr Wenger et al.

11:00–11:15 A Multilingual Approach to Identify and Classify Exceptional Measures against COVID-19
Georgios Tziafas et al.

11:15–11:30 Multi-granular Legal Topic Classification on Greek Legislation
Christos Papaloukas et al.

11:30–11:40 Machine Extraction of Tax Laws from Legislative Texts
Elliott Ash, Malka Guillot, and Luyang Han

11:40–11:50 jurBERT: A Romanian BERT Model for Legal Judgement Prediction
Mihai Masala et al.

11:50–12:00 JuriBERT: A Masked-Language Model Adaptation for French Legal Text
Stella Douka et al.

12:00–12:10 Few-shot and Zero-shot Approaches to Legal Text Classification: A Case Study in the Financial Sector
Rajdeep Sarkar et al.

12:10–1:00 **Lunch break**

1:00–1:45 **Invited Talk: Sylvie Delacroix**

2:00–2:10 A Free Format Legal Question Answering System
Soha Khazaeli et al.

2:10–2:20 Searching for Legal Documents at Paragraph Level: Automating Label Generation and Use of an Extended Attention Mask for Boosting Neural Models of Semantic Similarity
Li Tang and Simon Clematide

2:20–2:30 GerDaLIR: A German Dataset for Legal Information Retrieval
Marco Wrzalik and Dirk Krechel

2:45–3:00 SPaR.txt, a Cheap Shallow Parsing Approach for Regulatory Texts
Ruben Kruiper et al.

3:00–3:15 Capturing Logical Structure of Visually Structured Documents with Multimodal Transition Parser
Yuta Koreeda and Christopher Manning

3:15–3:30 Legal Terminology Extraction with the Termolator
Nhi Pham, Lachlan Pham, and Adam L. Meyers

3:30–3:45 Supervised Identification of Participant Slots in Contracts
Dan Simonson

4:00–4:10 Named Entity Recognition in Historic Legal Text: A Transformer and State Machine Ensemble Method
Fernando Trias et al.

4:25–4:40 Summarization of German Court Rulings
Ingo Glaser, Sebastian Moser, and Florian Matthes

4:55–5:10 Learning from Limited Labels for Long Legal Dialogue
Jenny Hong, Derek Chong, and Christopher Manning

5:10–5:20 Automating Claim Construction in Patent Applications: The CMUmine Dataset
Ozan Tonguz et al.

5:20–5:30 Effectively Leveraging BERT for Legal Document Classification
Nut Limsopatham

5:30–5:45 Semi-automatic Triage of Requests for Free Legal Assistance
Meladel Mistica et al.

5:45–6:00 Automatic Resolution of Domain Name Disputes
Wayan Oger Vihikan et al.

Workshop 18: SustaiNLP 2021: The Second Workshop on Simple and Efficient Natural Language Processing

Organizers: *Angela Fan et al.*

Wednesday, November 10, 2021

- Low Resource Quadratic Forms for Knowledge Graph Embeddings
  Zachary Zhou et al.
- Evaluating the carbon footprint of NLP methods: a survey and analysis of existing tools
  Nesrine bannour et al.
- Limitations of Knowledge Distillation for Zero-shot Transfer Learning
  Saleh Soltan, Haidar Khan, and Wael Hamza
- Countering the Influence of Essay Length in Neural Essay Scoring
  Sungho Jeon and Michael Strube
- Memory-efficient Transformers via Top-k Attention
  Ankit Gupta et al.
- BioCopy: A Plug-And-Play Span Copy Mechanism in Seq2Seq Models
  Yi Liu et al.
- Combining Lexical and Dense Retrieval for Computationally Efficient Multi-hop Question Answering
  Georgios Sidiropoulos et al.
- Learning to Rank in the Age of Muppets: Effectiveness—Efficiency Tradeoffs in Multi-Stage Ranking
  Yue Zhang et al.
- Improving Synonym Recommendation Using Sentence Context
  Maria Glenski et al.
- Semantic Categorization of Social Knowledge for Commonsense Question Answering
  Gengyu Wang et al.
- Speeding Up Transformer Training By Using Dataset Subsampling - An Exploratory Analysis
  Lovre Torbarina et al.
- Length-Adaptive Transformer: Train Once with Length Drop, Use Anytime with Search
  Gyuwan Kim and Kyunghyun Cho
- Hyperparameter Power Impact in Transformer Language Model Training
  Lucas Høyberg Puvis de Chavannes et al.
- Distiller: A Systematic Study of Model Distillation Methods in Natural Language Processing
  Haoyu He et al.
- Simple and Efficient ways to Improve REALM
  Vidhisha Balachandran et al.
- Shrinking Bigfoot: Reducing wav2vec 2.0 footprint
  Zilun Peng et al.
- On the Role of Corpus Ordering in Language Modeling
  Ameeta Agrawal et al.
- Efficient Domain Adaptation of Language Models via Adaptive Tokenization
  Vin Sachidananda, Jason Kessler, and Yi-An Lai
- Unsupervised Contextualized Document Representation
  Ankur Gupta and Vivek Gupta
- Logistic Regression Trained on Learner Data Outperformed Neural Language Models in Unsupervised Automatic Readability Assessment
  Yo Ehara

Workshop 20: Fourth Workshop on Fact Extraction and VERification (FEVER)

Organizers: *Rami Aly et al.*

Wednesday, November 10, 2021

9:00–9:05 **Opening Remarks**

9:05–10:20 **Keynote Talks**

9:05–9:35 **Keynote Talk (Preslav Nakov)**

9:35–10:05 **Keynote Talk (Maria Liakata)**

10:05–10:20 **QA**

10:20–11:00 **Research Talks**

10:20–10:35 Evidence Selection as a Token-Level Prediction Task
Dominik Stammbach

10:35–10:50 Graph Reasoning with Context-Aware Linearization for Interpretable Fact Extraction and Verification
Neema Kotonya et al.

10:50–11:00 **QA**

11:00–11:15 **Break**

11:15–12:30 **Keynote Talks**

11:15–11:45 **Keynote Talk (Mirella Lapata)**

11:45–12:15 **Keynote Talk (Pasquale Minervini)**

12:15–12:30 **QA**

12:30–1:30 **Lunch**

1:30–2:45 **Keynote and Shared Task**

1:30–2:00 **Keynote Talk (Mohit Bansal)**

2:00–2:10 **The Fact Extraction and VERification Over Unstructured and Structured information (FEVEROUS) Shared Task**

2:10–2:20 FaBULOUS: Fact-checking Based on Understanding of Language Over Unstructured and Structured information
Mostafa BOUZIANE et al.

2:20–2:30 Team Papelo at FEVEROUS: Multi-hop Evidence Pursuit
Christopher Malon

2:30–2:45 **QA**

2:45–4:15 **Virtual Poster Session**

- Modeling Entity Knowledge for Fact Verification
Yang Liu, Chenguang Zhu, and Michael Zeng
- Verdict Inference with Claim and Retrieved Elements Using RoBERTa
In-Zu Gi, Ting-Yu Fang, and Richard Tzong-Han Tsai
- Stance Detection in German News Articles
Laura Mascarell et al.
- FANG-COVID: A New Large-Scale Benchmark Dataset for Fake News Detection in German
Justus Mattern et al.
- Combining sentence and table evidence to predict veracity of factual claims using TaPaS and RoBERTa
Martin Funkquist
- Automatic Fact-Checking with Document-level Annotations using BERT and Multiple Instance Learning
Aalok Sathe and Joonsuk Park
- Neural Re-rankers for Evidence Retrieval in the FEVEROUS Task
Mohammed Saeed et al.

- A Fact Checking and Verification System for FEVEROUS Using a Zero-Shot Learning Approach
  Orkun Temiz et al.

4:15–5:30 **Keynote Talks**

4:15–4:45 **Keynote Talk (Steven Novella)**

4:45–5:15 **Keynote Talk (Brendan Nyhan)**

5:15–5:30 **QA**

5:30–5:45 **Closing Remarks**

Workshop 21: Third Workshop on New Frontiers in Summarization (NewSum)

Organizers: *Lu Wang et al.*

9:00–10:30 **Morning Session I**

9:00–9:10 **Open remarks (NewSum Organizers)**

9:10–10:00 **Keynote I (Sashi Narayan (Google))**

10:00–10:10 Sentence-level Planning for Especially Abstractive Summarization
Andreas Marfurt and James Henderson

10:10–10:20 Template-aware Attention Model for Earnings Call Report Generation
Yangchen Huang, Prashant K. Dhingra, and Seyed Danial Mohseni Taheri

10:20–10:25 Knowledge and Keywords Augmented Abstractive Sentence Summarization
Shuo Guan

10:25–10:30 Rewards with Negative Examples for Reinforced Topic-Focused Abstractive Summarization
Khalil Mrini, Can Liu, and Markus Dreyer

10:30–11:00 **Coffee break I**

11:00–12:00 **Morning session II**

11:00–11:50 **Keynote II (Sebastian Gehrmann (Google))**

11:50–12:00 A Novel Wikipedia based Dataset for Monolingual and Cross-Lingual Summarization
Mehwish Fatima and Michael Strube

12:00–1:00 **Lunch break**

1:00–2:30 **Afternoon session I**

1:00–1:50 **Keynote III (Asli Celikyilmaz (Facebook AI Research))**

1:50–1:55 Evaluation of Summarization Systems across Gender, Age, and Race
Anna Jørgensen and Anders Søgaard

1:55–2:00 Evaluation of Abstractive Summarisation Models with Machine Translation in Deliberative Processes
Miguel Arana-Catania et al.

2:00–2:10 Capturing Speaker Incorrectness: Speaker-Focused Post-Correction for Abstractive Dialogue Summarization
Dongyub Lee et al.

2:10–2:20 Measuring Similarity of Opinion-bearing Sentences
Wenyi Tay et al.

2:20–2:30 EASE: Extractive-Abstractive Summarization End-to-End using the Information Bottleneck Principle
Haoran Li et al.

2:30–3:00 **Coffee break**

3:00–3:35 **Afternoon session II**

3:00–3:10 Context or No Context? A preliminary exploration of human-in-the-loop approach for Incremental Temporal Summarization in meetings
Nicole Beckage et al.

3:10–3:20 Are We Summarizing the Right Way? A Survey of Dialogue Summarization Data Sets
Don Tuggener et al.

3:20–3:30 Modeling Endorsement for Multi-Document Abstractive Summarization
Logan Lebanoff et al.

3:30–3:35 SUBSUME: A Dataset for Subjective Summary Extraction from Wikipedia Documents
Nishant Yadav et al.

3:35–4:15 **EMNLP Finding papers - Summarization**

4:15–4:45 **Coffee break**

4:45–6:00 **Afternoon session III**

4:45–4:55 TLDR9+: A Large Scale Resource for Extreme Summarization of Social Media Posts
Sajad Sotudeh et al.

4:55–5:00 A New Dataset and Efficient Baselines for Document-level Text Simplification in German
Annette Rios et al.

5:00–6:00 **Mentoring Program**

Workshop 1: 3rd Workshop on Economics and Natural Language Processing (ECONLP)

Organizers: *Udo Hahn, Veronique Hoste, and Amanda Stent*

Thursday, November 11, 2021

9:00–9:30 **Opening remarks and status report on economics NLP (Udo Hahn)**

9:30–10:40 **Session 1: datasets for economic NLP**

9:30–10:00 A Fine-Grained Annotated Corpus for Target-Based Opinion Analysis of Economic and Financial Narratives
Jiahui Hu and Patrick Paroubek

10:00–10:20 EDGAR-CORPUS: Billions of Tokens Make The World Go Round
Lefteris Loukas et al.

10:20–10:40 The Global Banking Standards QA Dataset (GBS-QA)
Kyunghwan Sohn, Sunjae Kwon, and Jaesik Choi

10:40–10:50 **Short break**

10:50–12:00 **Session 2: transformer-based methodologies in economic NLP**

10:50–11:20 Corporate Bankruptcy Prediction with BERT Model
Alex Kim and Sangwon Yoon

11:20–11:40 Is Domain Adaptation Worth Your Investment? Comparing BERT and FinBERT on Financial Tasks
Bo Peng et al.

11:40–12:00 From Stock Prediction to Financial Relevance: Repurposing Attention Weights to Assess News Relevance Without Manual Annotations
Luciano Del Corro and Johannes Hoffart

12:00–12:45 **Keynote: (Gerard Hoberg)**

12:45–2:00 **Lunch Break**

2:00–3:10 **Session 3: applications in economic NLP**

2:00–2:30 Extracting Economic Signals from Central Bank Speeches
Maximilian Ahrens and Michael McMahon

2:30–2:50 Privacy enabled Financial Text Classification using Differential Privacy and Federated Learning
Priyam Basu et al.

2:50–3:10 Using Word Embedding to Reveal Monetary Policy Explanation Changes
Akira Matsui, Xiang Ren, and Emilio Ferrara

3:10–3:40 **Coffee Break**

3:40–5:00 **Session 4: applications in economic NLP**

3:40–4:10 Effective Use of Graph Convolution Network and Contextual Sub-Tree for Commodity News Event Extraction
Meisin Lee, Lay-Ki Soon, and Eu-Gene Siew

4:10–4:40 Cryptocurrency Day Trading and Framing Prediction in Microblog Discourse
Anna Paula Pawlicka Maule and Kristen Johnson

4:40–5:00 To What Extent Can English-as-a-Second Language Learners Read Economic News Texts?
Yo Ehara

5:00–5:45 **Discussion Round: Planning for an ECONLP Challenge Competition**

5:45–6:00 **Concluding remarks**

Workshop 2: LAW-DMR: The 15th Linguistic Annotation 3rd Designing Meaning Representations Joint Workshop

Organizers: *Claire Cardie et al.*

Thursday, November 11, 2021

8:45–9:00 **Introduction**

9:00–11:00 **Session A: Oral presentations:**

9:00–10:00 **Invited talk by Stephan Oepen: Linguistic Structure beyond Strings and Trees**

10:00–10:15 Zero-shot cross-lingual Meaning Representation Transfer: Annotation of Hungarian using the Prague Functional Generative Description
Attila Novák, Borbála Novák, and Csilla Novák

10:15–10:30 Theoretical and Practical Issues in the Semantic Annotation of Four Indigenous Languages
Jens E. L. Van Gysel et al.

10:30–10:45 Representing Implicit Positive Meaning of Negated Statements in AMR
Katharina Stein and Lucia Donatelli

10:45–11:00 AutoAspect: Automatic Annotation of Tense and Aspect for Uniform Meaning Representations
Daniel Chen, Martha Palmer, and Meagan Vigus

11:30–1:30 **Session B: Poster presentations:**

11:30–12:00 **Overview of poster presentations**

12:00–12:45 Can predicate-argument relationships be extracted from UD trees?
Adam Ek, Jean-Philippe Bernardy, and Stergios Chatzikyriakidis

12:00–12:45 Classifying Divergences in Cross-lingual AMR Pairs
Shira Wein and Nathan Schneider

12:00–12:45 A Linguistic Annotation Framework to Study Interactions in Multilingual Healthcare Conversational Forums
Ishani Mondal et al.

12:00–12:45 Sister Help: Data Augmentation for Frame-Semantic Role Labeling
Ayush Pancholy, Miriam R L Petruck, and Swabha Swayamdipta

12:00–12:45 A Corpus Study of Creating Rule-Based Enhanced Universal Dependencies for German
Teresa Bürkle, Stefan Grünewald, and Annemarie Friedrich

12:00–12:45 Subcategorizing Adverbials in Universal Conceptual Cognitive Annotation
Zhuxin Wang, Jakob Prange, and Nathan Schneider

12:00–12:45 Simplifying annotation of intersections in time normalization annotation: exploring syntactic and semantic validation
Peiwen Su and Steven Bethard

12:00–12:45 Overcoming the challenges in morphological annotation of Turkish in universal dependencies framework
Talha Bedir et al.

12:00–12:45 Automatic Entity State Annotation using the VerbNet Semantic Parser
Ghazaleh Kazeminejad et al.

12:00–12:45 On Releasing Annotator-Level Labels and Information in Datasets
Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz

12:45–1:30 **Panel discussion: what's the future of linguistic annotation?**

2:00–4:00 **Session C: Oral presentations:**

2:00–3:00 **Invited talk by Joyce Chai: Intuitive Physics in Commonsense Language Understanding**

3:00–3:15 Increasing Sentence-Level Comprehension Through Text Classification of Epistemic Functions
Maria Berger and Elizabeth Goldstein

3:15–3:30 Towards a Methodology Supporting Semiautomatic Annotation of HeadMovements in Video-recorded Conversations
Patrizia Paggio et al.

3:30–3:45 Intensionalizing Abstract Meaning Representations: Non-Veridicality and Scope
Gregor Williamson, Patrick Elliott, and Yuxin Ji

3:45–4:00 WikiGUM: Exhaustive Entity Linking for Wikification in 12 Genres
Jessica Lin and Amir Zeldes

4:00–4:15 **Concluding remarks**

Workshop 7: 7th Workshop on Noisy User-generated Text (W-NUT 2021)

Organizers: *Wei Xu et al.*

- Text Simplification for Comprehension-based Question-Answering
  TANVI DADU et al.
- Finding the needle in a haystack: Extraction of Informative COVID-19 Danish Tweets
  Benjamin Olsen and Barbara Plank
- Detecting Depression in Thai Blog Posts: a Dataset and a Baseline
  Mika Hämäläinen et al.
- Keyphrase Extraction with Incomplete Annotated Training Data
  Yanfei Lei et al.
- Fine-grained Temporal Relation Extraction with Ordered-Neuron LSTM and Graph Convolutional Networks
  Minh Tran Phu, Minh Van Nguyen, and Thien Huu Nguyen
- Does It Happen? Multi-hop Path Structures for Event Factuality Prediction with Graph Transformer Networks
  Duong Le and Thien Huu Nguyen
- Google-trickers, Yaminjeongeum, and Leetspeak: An Empirical Taxonomy for Intentionally Noisy User-Generated Text
  Won Ik Cho and Soomin Kim
- Description-based Label Attention Classifier for Explainable ICD-9 Classification
  Malte Feucht et al.
- A Text Editing Approach to Joint Japanese Word Segmentation, POS Tagging, and Lexical Normalization
  Shohei Higashiyama et al.
- Intrinsic evaluation of language models for code-switching
  Sik Feng Cheong, Hai Leong Chieu, and Jing Lim
- Can images help recognize entities? A study of the role of images for Multimodal NER
  Shuguang Chen et al.
- Perceived and Intended Sarcasm Detection with Graph Attention Networks
  Joan Plepi and Lucie Flek
- Hierarchical Character Tagger for Short Text Spelling Error Correction
  Mengyi Gao, Canran Xu, and Peng Shi
- Common Sense Bias in Semantic Role Labeling
  Heather Lent and Anders Søgaard
- PoliWAM: An Exploration of a Large Scale Corpus of Political Discussions on WhatsApp Messenger
  Vivek Srivastava and Mayank Singh
- ParsTwiNER: A Corpus for Named Entity Recognition at Informal Persian
  MohammadMahdi Aghajani, AliAkbar Badri, and Hamid Beigy
- DreamDrug - A crowdsourced NER dataset for detecting drugs in darknet markets
  Johannes Bogensperger et al.
- Comparing Grammatical Theories of Code-Mixing
  Adithya Pratapa and Monojit Choudhury
- Improving Punctuation Restoration for Speech Transcripts via External Data
  Xue-Yong Fu et al.
- Learning to Rank Question Answer Pairs with Bilateral Contrastive Data Augmentation
  Yang Deng, Wenxuan Zhang, and Wai Lam
- Mitigation of Diachronic Bias in Fake News Detection Dataset
  Taichi Murayama, Shoko Wakamiya, and Eiji ARAMAKI
- Understanding the Impact of UGC Specificities on Translation Quality
  José Carlos Rosales Núñez, Djamé Seddah, and Guillaume Wisniewski
- Noisy UGC Translation at the Character Level: Revisiting Open-Vocabulary Capabilities and Robustness of Char-Based Models
  José Carlos Rosales Núñez, Guillaume Wisniewski, and Djamé Seddah
- Changes in Twitter geolocations: Insights and suggestions for future usage
  Anna Kruspe et al.
- ConQuest: Contextual Question Paraphrasing through Answer-Aware Synthetic Question Generation
  Mostafa Mirshekari, Jing Gu, and Aaron Sisto

- NADE: A Benchmark for Robust Adverse Drug Events Extraction in Face of Negations
  Simone Scaboro et al.
- SpanAlign: Efficient Sequence Tagging Annotation Projection into Translated Data applied to Cross-Lingual Opinion Mining
  Léo Jacqmin et al.
- A Novel Framework for Detecting Important Subevents from Crisis Events via Dynamic Semantic Graphs
  Evangelia Spiliopoulou et al.
- Synthetic Data Generation and Multi-Task Learning for Extracting Temporal Information from Health-Related Narrative Text
  Heereen Shim et al.
- Neural-based RST Parsing And Analysis In Persuasive Discourse
  Jinfen Li and Lu Xiao
- BART for Post-Correction of OCR Newspaper Text
  Elizabeth Soper, Stanley Fujimoto, and Yen-Yun Yu
- Coping with Noisy Training Data Labels in Paraphrase Detection
  Teemu Vahtola et al.
- Knowledge Distillation with Noisy Labels for Natural Language Understanding
  Shivendra Bhardwaj et al.
- Integrating Transformers and Knowledge Graphs for Twitter Stance Detection
  Thomas Clark et al.
- Detecting Cross-Geographic Biases in Toxicity Modeling on Social Media
  Sayan Ghosh et al.
- Detection of Puffery on the English Wikipedia
  Amanda Bertsch and Steven Bethard
- Robustness and Sensitivity of BERT Models Predicting Alzheimer's Disease from Text
  Jekaterina Novikova
- Understanding Model Robustness to User-generated Noisy Texts
  Jakub Náplava et al.
- CIDEr-R: Robust Consensus-based Image Description Evaluation
  Gabriel Oliveira dos Santos, Esther Luna Colombini, and Sandra Avila
- Improved Named Entity Recognition for Noisy Call Center Transcripts
  Sam Davidson et al.
- Contrapositive Local Class Inference
  Omid Kashefi and Rebecca Hwa
- Improved Multilingual Language Model Pretraining for Social Media Text via Translation Pair Prediction
  Shubhanshu Mishra and Aria Haghighi
- Co-training for Commit Classification
  Jian Yi David Lee and Hai Leong Chieu
- Study of Manifestation of Civil Unrest on Twitter
  Abhinav Chinta et al.
- The Korean Morphologically Tight-Fitting Tokenizer for Noisy User-Generated Texts
  Sangah Lee and Hyopil Shin
- Character Transformations for Non-Autoregressive GEC Tagging
  Milan Straka, Jakub Náplava, and Jana Straková
- Can Character-based Language Models Improve Downstream Task Performances In Low-Resource And Noisy Language Scenarios?
  Arij riabi, Benoît Sagot, and Djamé Seddah
- "Something Something Hota Hai!" An Explainable Approach towards Sentiment Analysis on Indian Code-Mixed Data
  Aman Priyanshu et al.
- BERTweetFR : Domain Adaptation of Pre-Trained Language Models for French Tweets
  Yanzhu Guo et al.
- To What Extent Does Lexical Normalization Help English-as-a-Second Language Learners to Read Noisy English Texts?
  Yo Ehara

Workshop 8: The 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL)

Organizers: *Stefania Degaetano-Ortlieb et al.*

- The Early Modern Dutch Mediascape. Detecting Media Mentions in Chronicles Using Word Embeddings and CRF
  Alie Lassche and Roser Morante
- FrameNet-like Annotation of Olfactory Information in Texts
  Sara Tonelli and Stefano Menini
- Batavia asked for advice. Pretrained language models for Named Entity Recognition in historical texts.
  Sophie I. Arnoult, Lodewijk Petram, and Piek Vossen
- Quantifying Contextual Aspects of Inter-annotator Agreement in Intertextuality Research
  Enrique Manjavacas Arevalo, Laurence Mellerin, and Mike Kestemont
- The Multilingual Corpus of Survey Questionnaires Query Interface
  Danielly Sorato and Diana Zavala-Rojas
- The FairyNet Corpus - Character Networks for German Fairy Tales
  David Schmidt et al.
- End-to-end style-conditioned poetry generation: What does it take to learn from examples alone?
  Jörg Wöckener et al.
- Emotion Classification in German Plays with Transformer-based Language Models Pretrained on Historical and Contemporary Language
  Thomas Schmidt, Katrin Dennerlein, and Christian Wolff
- Automating the Detection of Poetic Features: The Limerick as Model Organism
  Almas Abdibayev et al.
- Unsupervised Adverbial Identification in Modern Chinese Literature
  Wenxiu Xie et al.
- Data-Driven Detection of General Chiasmi Using Lexical and Semantic Features
  Felix Schneider et al.
- Translationese in Russian Literary Texts
  Maria Kunilovskaya, Ekaterina Lapshinova-Koltunski, and Ruslan Mitkov
- BAHP: Benchmark of Assessing Word Embeddings in Historical Portuguese
  Zuoyu Tian et al.
- The diffusion of scientific terms – tracing individuals' influence in the history of science for English
  Yuri Bizzoni et al.
- A Pilot Study for BERT Language Modelling and Morphological Analysis for Ancient and Medieval Greek
  Pranaydeep Singh, Gorik Rutten, and Els Lefever
- Zero-Shot Information Extraction to Enhance a Knowledge Graph Describing Silk Textiles
  Thomas Schleider and Raphael Troncy
- 'Tecnologica cosa': Modeling Storyteller Personalities in Boccaccio's 'Decameron'
  A. Cooper et al.
- WMDecompose: A Framework for Leveraging the Interpretable Properties of Word Mover's Distance in Sociocultural Analysis
  Mikael Brunila and Jack LaViolette
- Period Classification in Chinese Historical Texts
  Zuoyu Tian and Sandra Kübler
- A Mixed-Methods Analysis of Western and Hong Kong—based Reporting on the 2019—2020 Protests
  Arya D. McCarthy, James Scharf, and Giovanna Maria Dora Dore
- Stylometric Literariness Classification: the Case of Stephen King
  Andreas van Cranenburgh and Erik Ketzan

Workshop 13: The 1st Workshop on Multi-lingual Representation Learning (MRL)

Organizers: *Duygu Ataman et al.*

Thursday, November 11, 2021

10:00–10:30 Language Models are Few-shot Multilingual Learners
Genta Indra Winata et al.

10:30–11:00 Learning Contextualised Cross-lingual Word Embeddings and Alignments for Extremely Low-Resource Languages Using Parallel Corpora
Takashi Wada et al.

11:00–12:00

- Clustering Monolingual Vocabularies to Improve Cross-Lingual Generalization
Riccardo Bassani, Anders Søgaard, and Tejaswini Deoskar
- Do not neglect related languages: The case of low-resource Occitan cross-lingual word embeddings
Lisa Woller, Viktor Hangya, and Alexander Fraser
- Specializing Multilingual Language Models: An Empirical Study
Ethan C. Chau and Noah A. Smith
- Learning Cross-lingual Representations for Event Coreference Resolution with Multi-view Alignment and Optimal Transport
Duy Phung et al.
- Multilingual and Multilabel Emotion Recognition using Virtual Adversarial Training
Vikram Gupta
- Analyzing the Effects of Reasoning Types on Cross-Lingual Transfer Performance
Karthikeyan K et al.
- Identifying the Importance of Content Overlap for Better Cross-lingual Embedding Mappings
Réka Cserháti and Gábor Berend
- On the Cross-lingual Transferability of Contextualized Sense Embeddings
Kiamehr Rezaee et al.
- Small Data? No Problem! Exploring the Viability of Pretrained Multilingual Language Models for Low-resourced Languages
Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin

1:45–2:15 Mr. TyDi: A Multi-lingual Benchmark for Dense Retrieval
Xinyu Zhang et al.

2:15–2:45 VisualSem: a high-quality knowledge graph for vision and language
Houda Alberts et al.

2:45–3:45

- Vykarana: A Colorless Green Benchmark for Syntactic Evaluation in Indic Languages
Rajaswa Patil et al.
- Improving the Diversity of Unsupervised Paraphrasing with Embedding Outputs
Monisha Jegadeesan et al.
- The Effectiveness of Intermediate-Task Training for Code-Switched Natural Language Understanding
Archiki Prasad et al.
- Shaking Syntactic Trees on the Sesame Street: Multilingual Probing with Controllable Perturbations
Ekaterina Taktasheva, Vladislav Mikhailov, and Ekaterina Artemova
- Multilingual Code-Switching for Zero-Shot Cross-Lingual Intent Prediction and Slot Filling
Jitin Krishnan et al.
- Analysis of Zero-Shot Crosslingual Learning between English and Korean for Named Entity Recognition
Jongin Kim et al.
- Regularising Fisher Information Improves Cross-lingual Generalisation
Asa Cooper Stickland and Iain Murray
- DMix: Distance Constrained Interpolative Mixup
Ramit Sawhney et al.
- Sequence Mixup for Zero-Shot Cross-Lingual Part-Of-Speech Tagging
Megh Thakkar et al.

- Well-Defined Morphology is Sentence-Level Morphology
  Omer Goldman and Reut Tsarfaty
- Cross-Lingual Training of Dense Retrievers for Document Retrieval
  Peng Shi et al.

Workshop 14: The Fifth Widening NLP Workshop (WiNLP 2021)

Organizers: *Khyathi Raghavi Chandu et al.*

Thursday, November 11, 2021

8:00–9:00 **Poster Session A**

9:00–9:10 **Opening Remarks**

9:10–9:50 **Joint Keynote speaker with Queer In AI: Dr. Jasmijn Bastings (English with Spanish translation)**

9:50–10:30 **Keynote speaker 2: Prof. Manuel Montes (Spanish with English translation)**

10:30–11:00 **Coffee Break**

11:00–12:00 **Tutorial: How to write and respond to a review**

12:00–1:00 **Lunch**

1:00–2:20 **Panel: The Peer Review Process and Widening NLP**

2:20–2:45 **Coffee break**

2:45–3:25 **Keynote speaker 3: Dra. Adriana Lorena Iiguez Carrillo**

3:25–4:25 **Poster Session B**

- OkwuGbé: End-to-End Speech Recognition for Fon and Igbo
  Bonaventure F. P. Dossou and Chris Chinenye Emezue
- TEET! Tunisian Dataset for Toxic Speech Detection
  Slim Gharbi et al.
- Ara-Women-Hate: The first Arabic Hate Speech corpus regarding Women
  Imane Guellil et al.
- Multi-label Emotion Classification on Code-Mixed Text: Data and Methods
  Iqra Ameer et al.
- Behavioral Testing of Knowledge Graph Embedding Models for Link Prediction
  Wiem Ben Rim et al.
- Developing Language Technology and NLP tools for endangered languages: Torwali
  Naeem Uddin Hadi
- How to Make Virtual Conferences Queer-Friendly: A Guide
  Organizers of QueerInAI et al.
- Neutralizing Gender Bias in Neural Machine Translation by Introducing Linguistic Knowledge
  Ksenia Kharitonova et al.
- Developing Keyboards for the Endangered Livonian Language
  Mika Hämäläinen and Khalid Alnajjar
- Nuanced Queerphobic Bias in Popular Sentiment Analysis Tools: A Data Set and Evaluation
  Eddie Ungless, Björn Ross, and Vaishak Belle
- Coral: An Approach for Conversational Agents in Mental Health Applications
  Harsh Sakhrani, Saloni Parekh, and Shubham Mahajan
- EM ALBERT: a step towards equipping Manipuri for NLP
  Rudali Huidrom and Yves Lepage
- Elementary-Level Math Word Problem Generation using Pre-Trained Transformers
  Kashyapa Niyarepola, Savindu Ekanayake, and Dineth Athapaththu
- Towards the Early Detection of Child Predators in Chat Rooms: A BERT-based Approach
  Sinchana Kumbale and Smriti Singh
- One-Shot Lexicon Learning for Low-Resource Machine Translation
  Anjali Kantharuban and Jacob Andreas
- Sinhala-English Code-mixed and Code-switched Data Classification
  Janani Sumanapala, Himashi Rathnayake, and Raveesha Rukshani
- ”I don’t know who she is”: Discourse and Knowledge Driven Coreference Resolution
  Angela Ramirez et al.

- Idiom Extraction Method with Fine-tuning of Pre-trained TransformersModel for Named Entity Recognition
  Nao Yamato
- Occupational Gender stereotypes in Indic Languages
  Neeraja Kirtane and Tanvi Anand
- #WhyDidTheyStay: An NLP-driven approach to analyzing the factors that affect domestic violence victims
  Marthala Kavya and Smriti Singh
- Exploring Transfer Learning Pathways for Neural Machine Back Translation of Eskimo-Aleut, Chicham, and Classical Languages
  Aaron Serianni and Daniel Whitenack
- An interpretable representation that visually grounds dialog history
  Mauricio Mazuecos et al.
- Automated Template Paraphrasing for Conversational Assistants
  Liane Vogel and Lucie Flek
- Discovering changes in birthing narratives during COVID-19
  Daphna Spira et al.
- Explorations in Transfer Learning for OCR Post-Correction
  Lindia Tjuatja, Shruti Rijhwani, and Graham Neubig
- A Prototype Free/Open-Source Morphological Analyser and Generator for Sakha
  Sardana Ivanova, Francis Tyers, and Jonathan Washington
- Towards Text Simplification for Sinhala Language
  Rumesh Sirithunga, Lahiru De Silva, and Thamindu Aluthwala
- Natural language processing as a tool to identify the Reddit particularities of cancer survivors around the time of diagnosis and remission: A pilot study
  Ioana R. Podină et al.
- Identifying Significant Citations via Mining Paper Full-Text
  Muskaan Singh and Tirthankar Ghosal
- SciBERT-based Multitasking Deep Neural Architecture to identify Contribution Statements from Scientific articles
  Komal Gupta and Tirthankar Ghosal
- Characterizing Test Anxiety on Social Media
  Esha Julka et al.
- Sample Selection Guided by Domain and Task for Cross-Domain Targeted Sentiment Analysis
  Kasturi Bhattacharjee, Rashmi Gangadharaiah, and Smaranda Muresan
- Bengali Parallel Universal Dependency Treebank
  Pritha Majumdar
- The Development of Pre-processing Tools and Pre-trained Embedding Models for Amharic
  Tadesse Destaw, Abinew Ayele, and Seid Muhie Yimam
- Towards Syntax-Aware DialogueSummarization using Multi-task Learning
  Seolhwa Lee et al.
- Detoxifying Language Models with Proximal Policy Optimization
  Taaha Kazi
- Towards Personalized Descriptions of Scientific Concepts
  Sonia Murthy et al.
- Building Prosody Labeled Corpus in Hindi
  Esha Banerjee, Atul Kr. Ojha, and Girish Jha
- How Well Can an Agent Understand Different Accents?
  Divya Tadimeti, Kallirroi Georgila, and David Traum
- Monolingual Pre-Trained Language Models for Tigrinya
  Fitsum Gaim, Wonsuk Yang, and Jong C. Park
- ASQ: Automatically Generating Question-Answer Pairs using AMRs
  Geetanjali Rakshit and Jeffrey Flanigan
- Adverse Drug Reaction Classification of Tweets with Fusion of Text and Drug Representations
  Andrey Sakhovskiy and Elena Tutubalina
- Detecting Gender Bias using Explainability
  Gauri Gupta, Supriti Vijay, and Krithika Ramesh
- Leveraging ultradense embeddings to analyze gender-oriented extremist recruitment targeting
  Jatin Khilnani et al.

4:25–4:35 **Closing Remarks**

4:35–5:30 **Social Coffee**

Workshop 17: Workshop on Evaluations and Assessments of Neural Conversation Systems

Organizers: *Wei Wei et al.*

- Counterfactual Matters: Intrinsic Probing For Dialogue State Tracking
  Yi Huang et al.
- GCDF1: A Goal- and Context- Driven F-Score for Evaluating User Models
  Alexandru Coca, Bo-Hsiang Tseng, and Bill Byrne
- Unsupervised Testing of NLU models with Multiple Views
  Radhika Arava et al.
- A Comprehensive Assessment of Dialog Evaluation Metrics
  Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri

Workshop 19: Fourth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2021)

Organizers: *Maciej Ogrodniczuk et al.*

Wednesday, November 10, 2021

Session 1: Opening Remarks and System Talks (Part 1)

9:00–9:30 **Welcome and Task Overview**

9:30–9:45 **Neural Anaphora Resolution in Dialogue**

9:45–10:00 **Anaphora Resolution in Dialogue: Description of the DFKI-Talking Robots System for the CODI-CRAC 2021 Shared-Task**

10:00–10:30 **Coffee Break**

Session 2: System Talks (Part 2), Cross-Team Analyses, and Discussion

10:30–10:45 **The Pipeline Model for Resolution of Anaphoric Reference and Resolution of Entity Reference**

10:45–11:00 **An End-to-End Approach for Full Bridging Resolution**

11:00–11:15 **Adapted End-to-End Coreference Resolution System for Anaphoric Identities in Dialogues**

11:15–11:30 **Anaphora Resolution in Dialogue: Cross-Team Analysis of the DFKI-TalkingRobots Team Submissions for the CODI-CRAC 2021 Shared-Task**

11:30–11:45 **The CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis Resolution: A Cross-Team Analysis**

11:45–12:00 **Visioning Discussion and Next Steps**

Thursday, November 11, 2021

9:00–9:05 **Opening Remarks**

Paper Session 1

9:05–9:20 A Brief Survey and Comparative Study of Recent Development of Pronoun Coreference Resolution in English
Hongming Zhang, Xinran Zhao, and Yangqiu Song

9:20–9:35 Coreference Resolution for the Biomedical Domain: A Survey
Pengcheng Lu and Massimo Poesio

9:35–9:50 FantasyCoref: Coreference Resolution on Fantasy Literature Through Omniscient Writer's Point of View
Sooyoun Han et al.

9:50–10:00 **CDLM: Cross-Document Language Modeling**

10:00–10:15 DramaCoref: A Hybrid Coreference Resolution System for German Theater Plays
Janis Pagel and Nils Reiter

10:15–10:30 A Hybrid Rule-Based and Neural Coreference Resolution System with an Evaluation on Dutch Literature
Andreas van Cranenburgh et al.

Invited Talk

11:00–12:00 **Modeling informational relations across multiple texts Ido Dagan**

12:00–1:00 **Lunch Break**

Paper Session 2

1:00–1:10 Lazy Low-Resource Coreference Resolution: a Study on Leveraging Black-Box Translation Tools
Semere Kiros Bitew et al.

1:10–1:20 **Collecting a Large-Scale Gender Bias Dataset for Coreference Resolution and Machine Translation**

1:20–1:30 **Do UD Trees Match Mention Spans in Coreference Annotations?**

1:30–1:40 **End-to-end Neural Information Status Classification**

1:40–1:50 Resources and Evaluations for Danish Entity Resolution
Maria Barrett et al.

1:50–2:05 CoreLM: Coreference-aware Language Model Fine-Tuning
Nikolaos Stylianou and Ioannis Vlahavas

2:05–2:20 Data Augmentation Methods for Anaphoric Zero Pronouns
Abdulrahman Aloraini and Massimo Poesio

2:20–2:30 Exploring Pre-Trained Transformers and Bilingual Transfer Learning for Arabic Coreference Resolution
Bonan Min

2:30–2:45 **Mini Break**

Universal Anaphora Panel

2:55–3:05 **The Universal Anaphora Extension of the CONLL-U Markup Scheme**

3:05–3:15 **The Universal Anaphora Scorer**

3:15–3:25 **The CODI/CRAC Shared Task on Anaphora Resolution in Dialogue**

3:25–3:45 **Discussion**

Best Papers Session

3:45–4:00 Event and Entity Coreference using Trees to Encode Uncertainty in Joint Decisions
Nishant Yadav et al.

4:00–4:15 On Generalization in Coreference Resolution
Shubham Toshniwal et al.

4:15–4:45 **Coffee Break**

Paper Session 3

4:45–5:00 Improving Span Representation for Domain-adapted Coreference Resolution
Nupoor Gandhi, Anjalie Field, and Yulia Tsvetkov

5:00–5:15 Coreference by Appearance: Visually Grounded Event Coreference Resolution
Liming Wang et al.

5:15–5:30 Anatomy of OntoGUM—Adapting GUM to the OntoNotes Scheme to Evaluate Robustness of SOTA Coreference Algorithms
Yilun Zhu, Sameer Pradhan, and Amir Zeldes

5:30–5:40 Understanding Mention Detector-Linker Interaction in Neural Coreference Resolution
Zhaofeng Wu and Matt Gardner

5:40–5:50 **Coreference-aware Surprisal Predicts Brain Response**

5:50–6:00 **Closing Remarks**

Workshop 23: BlackboxNLP: Analyzing and interpreting neural networks for NLP

Organizers: *Dieuwke Hupkes et al.*

Thursday, November 11, 2021

Virtual Programme

2:00–2:15 **Opening Remarks**

2:15–3:00 **Invited Talk by Jelle Zuidema & Q&A**

3:15–4:00 **Oral Session 1**

4:30–6:00 **Poster Session 1**

6:15–7:00 **Oral Session 2**

7:30–8:00 **Invited Talk by Ana Marasović**

8:00–8:30 **Invited Talk by Sara Hooker**

8:30–8:45 **Closing**

Hybrid Programme

9:00–9:15 **Opening Remarks & Best Paper Award**

9:15–10:00 **Invited Talk by Jelle Zuidema & Q&A**

10:00–10:30 **Oral Session 3**

11:00–12:00 **Poster Session 2**

1:00–1:45 **Invited Talk by Sara Hooker & Q&A**

1:45–2:15 **Oral Session 4**

2:45–4:15 **Poster Session 3**

4:45–5:15 **Oral Session 5**

5:15–6:00 **Invited Talk by Ana Marasović & Q&A**

6:00–6:15 **Closing**

Oral Sessions

- To what extent do human explanations of model behavior align with actual model behavior?
  Grusha Prasad et al.
- Test Harder than You Train: Probing with Extrapolation Splits
  Jenny Kunz and Marco Kuhlmann
- Does External Knowledge Help Explainable Natural Language Inference? Automatic Evaluation vs. Human Ratings
  Hendrik Schuff et al.
- The Language Model Understood the Prompt was Ambiguous: Probing Syntactic Uncertainty Through Generation
  Laura Aina and Tal Linzen
- On the Limits of Minimal Pairs in Contrastive Evaluation
  Jannis Vamvas and Rico Sennrich
- What Models Know About Their Attackers: Deriving Attacker Information From Latent Representations
  Zhouhang Xie et al.

Poster Sessions

- ALL Dolphins Are Intelligent and SOME Are Friendly: Probing BERT for Nouns' Semantic Properties and their Prototypicality
  Marianna Apidianaki and Aina Garí Soler
- ProSPer: Probing Human and Neural Network Language Model Understanding of Spatial Perspective
  Tessa Masis and Carolyn Anderson

- Can Transformers Jump Around Right in Natural Language? Assessing Performance Transfer from SCAN
  Rahma Chaabouni, Roberto Dessì, and Eugene Kharitonov
- Transferring Knowledge from Vision to Language: How to Achieve it and how to Measure it?
  Tobias Norlund, Lovisa Hagström, and Richard Johansson
- Discrete representations in neural models of spoken language
  Bertrand Higy et al.
- Word Equations: Inherently Interpretable Sparse Word Embeddings through Sparse Coding
  Adly Templeton
- A howling success or a working sea? Testing what BERT knows about metaphors
  Paolo Pedinotti et al.
- How Length Prediction Influence the Performance of Non-Autoregressive Translation?
  Minghan Wang et al.
- On the Language-specificity of Multilingual BERT and the Impact of Fine-tuning
  Marc Tanti et al.
- Relating Neural Text Degeneration to Exposure Bias
  Ting-Rui Chiang and Yun-Nung Chen
- Efficient Explanations from Empirical Explainers
  Robert Schwarzenberg, Nils Feldhus, and Sebastian Möller
- Variation and generality in encoding of syntactic anomaly information in sentence embeddings
  Qinxuan Wu and Allyson Ettinger
- Enhancing Interpretable Clauses Semantically using Pretrained Word Representation
  Rohan Kumar Yadav et al.
- Analyzing BERT's Knowledge of Hypernymy via Prompting
  Michael Hanna and David Mareček
- An in-depth look at Euclidean disk embeddings for structure preserving parsing
  Federico Fancellu et al.
- Training Dynamic based data filtering may not work for NLP datasets
  Arka Talukdar et al.
- Multi-Layer Random Perturbation Training for improving Model Generalization Efficiently
  Lis Kanashiro Pereira, Yuki Taya, and Ichiro Kobayashi
- Screening Gender Transfer in Neural Machine Translation
  Guillaume Wisniewski et al.
- What BERT Based Language Model Learns in Spoken Transcripts: An Empirical Study
  Ayush Kumar, Mukuntha Narayanan Sundararaman, and Jithendra Vepa
- Assessing the Generalization Capacity of Pre-trained Language Models through Japanese Adversarial Natural Language Inference
  Hitomi Yanaka and Koji Mineshima
- Investigating Negation in Pre-trained Vision-and-language Models
  Radina Dobreva and Frank Keller
- Not all parameters are born equal: Attention is mostly what you need
  Nikolay Bogoychev
- Not All Models Localize Linguistic Knowledge in the Same Place: A Layer-wise Probing on BERToids' Representations
  Mohsen Fayyaz et al.
- Learning Mathematical Properties of Integers
  Maria Ryskina and Kevin Knight
- Probing Language Models for Understanding of Temporal Expressions
  Shivin Thukral, Kunal Kukreja, and Christian Kavouras
- How Familiar Does That Sound? Cross-Lingual Representational Similarity Analysis of Acoustic Word Embeddings
  Badr Abdullah et al.
- Perturbing Inputs for Fragile Interpretations in Deep Natural Language Processing
  Sanchit Sinha et al.
- An Investigation of Language Model Interpretability via Sentence Editing
  Samuel Stevens and Yu Su
- Interacting Knowledge Sources, Inspection and Analysis: Case-studies on Biomedical text processing
  Parsa Bagherzadeh and Sabine Bergler
- Attacks against Ranking Algorithms with Text Embeddings: A Case Study on Recruitment Algorithms
  Anahita Samadi and Shirin Nilizadeh

- Controlled tasks for model analysis: Retrieving discrete information from sequences
  Ionut-Teodor Sorodoc, Gemma Boleda, and Marco Baroni
- The Acceptability Delta Criterion: Testing Knowledge of Language using the Gradience of Sentence Acceptability
  Héctor Vázquez Martínez
- How Does BERT Rerank Passages? An Attribution Analysis with Information Bottlenecks
  Zhiying Jiang et al.
- Do Language Models Know the Way to Rome?
  Bastien Liétard, Mostafa Abdou, and Anders Søgaard
- Exploratory Model Analysis Using Data-Driven Neuron Representations
  Daisuke Oba, Naoki Yoshinaga, and Masashi Toyoda
- Fine-Tuned Transformers Show Clusters of Similar Representations Across Layers
  Jason Phang, Haokun Liu, and Samuel R. Bowman
- BERT Has Uncommon Sense: Similarity Ranking for Word Sense BERTology
  Luke Gessler and Nathan Schneider

Non-archival papers (posters)

8

Anti-harassment policy

The open exchange of ideas, the freedom of thought and expression, and respectful scientific debate are central to the aims and goals of the ACL. These require a community and an environment that recognizes the inherent worth of every person and group, that fosters dignity, understanding, and mutual respect, and that embraces diversity. For these reasons, ACL is dedicated to providing a harassment-free experience for all the members, as well as participants at our events and in our programs.

Harassment and hostile behavior are unwelcome at any ACL conference, associated event, or in ACL-affiliated on-line discussions. This includes: speech or behavior that intimidates, creates discomfort, or interferes with a person's participation or opportunity for participation in a conference or an event. We aim for ACL-related activities to be an environment where harassment in any form does not happen, including but not limited to: harassment based on race, gender, religion, age, color, appearance, national origin, ancestry, disability, sexual orientation, or gender identity. Harassment includes degrading verbal comments, deliberate intimidation, stalking, harassing photography or recording, inappropriate physical contact, and unwelcome sexual attention. The policy is not intended to inhibit challenging scientific debate, but rather to promote it through ensuring that all are welcome to participate in shared spirit of scientific inquiry. Vexatious complaints and willful misuse of this procedure will render the complainant subject to the same sanctions as a violation of the anti-harassment policy.

It is the responsibility of the community as a whole to promote an inclusive and positive environment for our scholarly activities. In addition, anyone who experiences harassment or hostile behavior may contact any current member of the ACL Executive Committee or contact Priscilla Rasmussen (acl@aclweb.org), who is usually available at the registration desk during ACL conferences. Members of the executive committee will be instructed to keep any such contact in strict confidence, and those who approach the committee will be consulted before any actions are taken.

Implementation

This policy should be posted prominently on all ACL conference and workshop webpages, with a notice of a list of people who can be contacted by community members with concerns or complaints, which will be forwarded to the Professional Conduct Committee for investigation.

Approved by ACL Executive Committee, 2016

Revised by ACL Executive Committee, July 15, 2018

The policy is also available from ACL's main page.

9

Local Guide

COVID-19 Tests

Transportation and Tours:

Global Incentive Management, a destination management company, is the official tour and transportation company for EMNLP 2021 to ensure you have a safe, stress-free start and end to your trip in Punta Cana. Avoid overpaying taxis and other travel pitfalls. With Global Incentive Management, you will experience comfortable, economical service in an air-conditioned vehicle, with our staff greeting you at the Punta Cana airport and the best provider to serve you and your guests. There will be a desk available onsite at the resort for limited hours where you can make tour and other plans. You can also find more about the fun excursions they offer and make your transportation and tour plans directly through Global Incentive Management at this site:

https://gimdmc.com/EMNLP/index.html

Childcare

At the present, only individual childcare can be offered. The Kids Club is still closed per local COVID-19 restrictions. Individual childcare costs are $20 USD for one hour and any additional hours are $15 USD per hour. 24 hours notice is required.

The Human Language Technology Center of Excellence at Johns Hopkins University was founded in 2007 to create next-generation algorithms for speech and language processing. We identify and create innovative technologies that could have significant impact on challenging real-world problems. Our research addresses key challenges in extracting information from massive sources of text and speech.

We're always looking for talented personnel to join our team! We offer opportunities for summer hires for our SCALE summer workshop as well as full-time researcher hires. We welcome all inquiries.

hltcoe-hiring@jhu.edu

EMNLP 2021 gratefully acknowledges the following sponsors for their support:

Diamond


Platinum

Gold

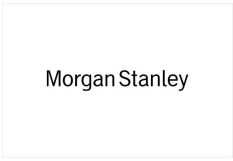
Silver

Bronze

Supporter

Diversity and Inclusion: Champion

Diversity and Inclusion: Ally



Diversity and Inclusion: Contributor