EMNLP 2021

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

The 2021 Conference on Empirical Methods in Natural Language Processing

CONFERENCE HANDBOOK

Meeting rooms

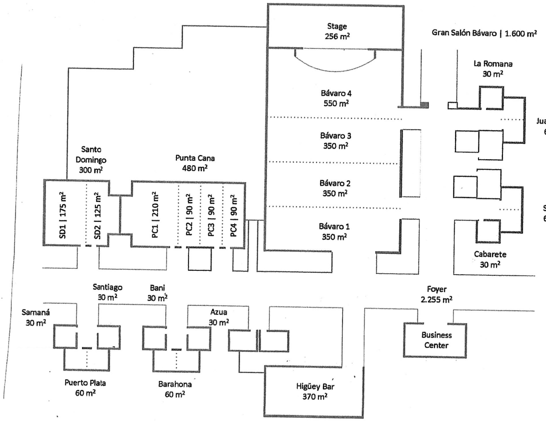
Registration: In front of Business Center

Main Conference:

Bavaro 1: Part of Plenary + Parallel
Bavaro 2: Part of Plenary + Parallel
Bavaro 3: Part of Plenary + Parallel
Punta Cana: Parallel
Santo Domingo: Parallel
Santiago and Samana: Mother's Rooms
Cabarete: Prayer Room
Juan Dolio: Quiet Room
All Posters and Demos for
Main Conference and Workshops:
Bavaro 4

All Tutorials and Workshops:

Using all rooms above plus most smaller
Rooms—See Room Assignment Handout



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

Table of Contents **i**

1 Conference Information **1**
Message from the General Chair . . . 1
Message from the Program Committee Co-Chairs . . . 3
Organizing Committee . . . 7
Program Committee . . . 8
Welcome Reception . . . 15

2 Tutorials: Wednesday, November 10 **17**
T1: Crowdsourcing Beyond Annotation:
Case Studies in Benchmark Data Collection . . . 18
T2: Financial Opinion Mining . . . 20
T3: Knowledge-Enriched Natural Language Generation . . . 22

3 Tutorials: Thursday, November 11 **25**
T4: Multi-Domain Multilingual Question Answering . . . 26
T5: Robustness and Adversarial Examples in Natural Language Processing . . . 27
T6: Syntax in End-to-End Natural Language Processing . . . 29

4 Main Conference: Sunday, November 7 **31**
Keynote Address: Ido Dagan . . . 32
Session 1 . . . 34
Session 2 . . . 48
Session 3 . . . 62
Session 4 . . . 75

5 Main Conference: Monday, November 8 **89**
Keynote Address: Evelina Fedorenko . . . 90
Session 5 . . . 91
Session 6 . . . 105
Session 7 . . . 119

Social Event 134

6 Main Conference: Tuesday, November 9 **135**
Keynote Address: Steven Bird 136
Session 8 137
Session 9 150

7 Workshops: Thursday–Friday, November 10–11 **161**
W22: The 25th Conference on Computational Natural Language Learning (CoNLL) . 162
W9: 8th International Workshop on Argument Mining 165
W16: 2nd Workshop on Computational Approaches to Discourse (CODI) 166
W3: 3rd Workshop on NLP for Conversational AI (NLP4ConvAI) 169
W6: CI+NLP: First Workshop on Causal Inference in NLP 171
W10: The Second Workshop on Insights from Negative Results in NLP 172
W12: The 3rd Workshop on Machine Reading for Question Answering 173
W15: The Natural Legal Language Processing Workshop 2021 (NLLP) 176
W18: SustaiNLP 2021: The Second Workshop on Simple and Efficient Natural Language Processing 177
W20: Fourth Workshop on Fact Extraction and VERification (FEVER) 178
W21: Third Workshop on New Frontiers in Summarization (NewSum) 179
W1: 3rd Workshop on Economics and Natural Language Processing (ECONLP) . . . 181
W7: 7th Workshop on Noisy User-generated Text (W-NUT 2021) 182
W8: The 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL) 184
W17: Workshop on Evaluations and Assessments of Neural Conversation Systems . . 185
W19: Fourth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2021) 186
W23: BlackboxNLP: Analyzing and interpreting neural networks for NLP 188

8 Anti-harassment policy **189**

Author Index **191**

9 Local Guide **191**

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. 841 submissions were accepted to the main conference. Among them, 315 were accepted as oral papers, and 526 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	402	124	526
Acceptance Rate (Main Conference)	25.6%	17.9%	23.4%
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 237 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 S(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 is 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;

- 237 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, 369 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 in a very tight schedule;

- 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 2021), 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 Glava?, 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

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

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 Research (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
Youzheng Wu, JD AI Research
Wei Wu, meituan
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

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

Efficient Methods for NLP

Jesse Dodge, (senior chair)
Gabriel Stanovsky, (senior chair)
Aleksandr Drozd, RIKEN Center for Computational Science
Adhi Kuncoro, University of Oxford and DeepMind
Emma Strubell, Carnegie Mellon University
Andreas Rückl, Amazon
Benoît Sagot, Inria
Jonathan Frankle, MIT

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, Peking University
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
Mo Yu, IBM Research
Dian Yu, Tencent AI Lab

Information Retrieval and Text Mining

Mark Sanderson, RMIT University (senior chair)
Andrew Yates, Max Planck Institute for Informatics (senior chair)
Simone Filice, amazon.com
Ahmed Hassan Awadallah, Microsoft Research
Evangelos Kanoulas, University of Amsterdam
Sarvnaz Karimi, amazon.com
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

Zack Lipton, Carnegie Mellon University (senior chair)
Francesca Toni, Imperial College London (senior chair)
Leila Arras, Fraunhofer Heinrich Hertz Institute
Jasmijn Bastings, Google
Grzegorz Chrupala, Tilburg University
Kevin Clark, Stanford University
Yonatan Belinkov, Technion
Sebastian Gehrmann, King's College London
Diewke 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
Oana Cocarascu, King's College London

Linguistic Theories, Cognitive Modeling and Psycholinguistics

Ryan Cotterell, ETH Zürich (senior chair)
Afra Alishahi, Tilburg University
Kyle Mahowald, University of California, Santa Barbara
Aida Nematzadeh, DeepMind
Adina Williams, Facebook, Inc.

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, ASAPP Inc
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, Toyota Technological Institute at Chicago

Machine Translation and Multilinguality
Alexandra Birch, University of Edinburgh (senior chair)
Yang Feng, Institute of Computing Technology (senior chair)
Veselin Stoyanov, Facebook (senior chair)
Boxing Chen, Alibaba
Colin Cherry, Google
Trevor Cohn, University of Melbourne
Marta Ruiz Costa-jussà, Universitat Politècnica de Catalunya
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
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 Sághdha, 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, University of Lisbon
Aston Zhang, AWS AI
Arkaitz Zubiaga, Queen Mary University of London

Phonology, Morphology and Word Segmentation
Xipeng Qiu, Queen Mary University of London (senior chair)
Baobao Chao, Institute of Computational Linguistic, Peking University
Ryan Cotterell, ETH Zürich
Hinrich Schütze, University of Munich
Hai Zhao, Shanghai Jiao Tong University
Tristan Naumann, Microsoft Research

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, Facebook AI Research
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)
Barbara Plank, IT University of Copenhagen
Ines Rehbein, University of Mannheim
Maja Popovic, ADAPT, Dublin City University
Gerasimos Lampouras, Huawei Noah's Ark Lab
Markus Freitag, Google Research
Simon Mille, Pompeu Fabra University
Ajay Nagesh, DiDi Labs
Gareth Jones, Dublin City University

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
Swabha Swayamdipta, University of Washington
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
Eric Cambria, Nanyang Technological University
Orphée De Clercq, LT3, Ghent University
Ivan Habernal, Technische Universität Darmstadt
Roman Klinger, University of Stuttgart
Anh Tuan Luu, NTU
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, Westlake 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 Ann Hendricks, DeepMind
Chiori Hori, Mitsubishi Electric Research Laboratories (MERL)
Douwe Kiela, Facebook
Florian Metze, Carnegie Mellon University
Tara Sainath, Google, Inc.
Radu Soricut, Google LLC
William Wang, Unversity 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
Anders Søgaard, University of Copenhagen

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

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.

16

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	**Welcome and conference logistics**	*Plaza Ballroom A, B, & C*
9:00 – 10:00	**INVITED TALK 1 (live) by Ido Dagan**	*Plaza Ballroom A, B, & C*

Session 1

10:30 – 12:00	Machine Translation and Multilinguality 1 *Plaza Ballroom A & B*	Summarization 1 *Plaza Ballroom D & E*	Information Extraction 1 *Plaza Ballroom F*	Sentiment Analysis, Stylistic Analysis, and Argument Mining 1	Computational Social Science and Cultural Analytics	Efficient Methods for NLP 1

Session 2

1:00 – 2:30	Machine Learning for NLP 1 *Plaza Ballroom A & B*	Generation 1 *Plaza Ballroom D & E*	Interpretability and Analysis of Models for NLP 1 *Plaza Ballroom F*	NLP Applications 1	Linguistic Theories, Cognitive Modeling and Psycholinguistics	Information Retrieval and Text Mining

Session 3

2:45 – 4:15	Machine Learning for NLP 2 *Plaza Ballroom A & B*	Dialogue and Interactive Systems 1 *Plaza Ballroom D & E*	Information Extraction 2 *Plaza Ballroom F*	Resources and Evaluation 1	Discourse and Pragmatics	Semantics 1

Session 4

4:45 – 6:15	Machine Learning for NLP 3 *Plaza Ballroom A & B*	Dialogue and Interactive Systems 2 *Plaza Ballroom D & E*	Information Extraction 3 *Plaza Ballroom F*	Ethics and NLP	Phonology, Morphology and Word Segmentation	Speech, Vision, Robotics, Multimodal Grounding 1

Keynote Address: Ido Dagan

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

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

Plaza Ballroom A, B, & C

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*
10:30	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.*	Unsupervised Keyphrase Extraction by Jointly Modeling Local and Global Context *Liang et al.*	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*
10:45	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*	Distantly Supervised Relation Extraction using Multi-Layer Revision Network and Confidence-based Multi-Instance Learning *Lin 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.*
11:00	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*	Logic-level Evidence Retrieval and Graph-based Verification Network for Table-based Fact Verification *Shi 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*

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*	
Cross Attention Augmented Transducer Networks for Simultaneous Translation *Liu et al.*	Decision-Focused Summarization *Hsu and Tan*	A Partition Filter Network for Joint Entity and Relation Extraction *Yan 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.*	11:15
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.*	TEBNER: Domain Specific Named Entity Recognition with Type Expanded Boundary-aware Network *Fang 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.*	11:30
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.*	TEBNER: Domain Specific Named Entity Recognition with Type Expanded Boundary-aware Network *Fang 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

Parallel Session 1

Session 1A: Machine Translation and Multilinguality 1

Plaza Ballroom A & B *Chair:*

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

Plaza Ballroom D & E *Chair:*

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

Plaza Ballroom F *Chair:*

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

Chair:

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

Chair:

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.

Session 1F: Efficient Methods for NLP 1

Chair:

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.

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.*	Truth-Conditional Captions for Time Series Data *Jhamtani and Berg-Kirkpatrick*	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.*	Injecting Entity Types into Entity-Guided Text Generation *Dong et al.*	Do Long-Range Language Models Actually Use Long-Range Context? *Sun et al.*	Voice Query Auto Completion *Tang 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
Active Learning by Acquiring Contrastive Examples *Margatina et al.*	Injecting Entity Types into Entity-Guided Text Generation *Dong et al.*	Do Long-Range Language Models Actually Use Long-Range Context? *Sun et al.*	Voice Query Auto Completion *Tang 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.*	2:00

Parallel Session 2

Session 2A: Machine Learning for NLP 1

Plaza Ballroom A & B *Chair:*

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.

Session 2B: Generation 1

Plaza Ballroom D & E *Chair:*

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.

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

Plaza Ballroom F *Chair:*

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.

Session 2D: NLP Applications 1

Chair:

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.

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

Chair:

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.

Session 2F: Information Retrieval and Text Mining

Chair:

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*
	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, Wooldridge, and Glick*	Stepmothers are mean and academics are pretentious: What do pretrained language models learn about you? *Choenni, Shutova, and Rooij*

Parallel Session 3

Session 3A: Machine Learning for NLP 2

Plaza Ballroom A & B *Chair:*

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.

Session 3B: Dialogue and Interactive Systems 1

Plaza Ballroom D & E *Chair:*

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

Plaza Ballroom F *Chair:*

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.

Session 3D: Resources and Evaluation 1

Chair:

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

Chair:

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 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.

Session 3F: Semantics 1

Chair:

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.

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*	
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.*	4:45
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*	5:00
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.*	5:15
Knowledge-Aware Meta-learning for Low-Resource Text Classification *Yao et al.*	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.*	Visual Goal-Step Inference using wikiHow *Yang et al.*	5:45

	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*
5:55	Knowledge-Aware Meta-learning for Low-Resource Text Classification *Yao et al.*	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.*	Visual Goal-Step Inference using wikiHow *Yang et al.*
6:05	Knowledge-Aware Meta-learning for Low-Resource Text Classification *Yao et al.*	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.*	Visual Goal-Step Inference using wikiHow *Yang et al.*

Parallel Session 4

Session 4A: Machine Learning for NLP 3

Plaza Ballroom A & B *Chair:*

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.

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

Plaza Ballroom D & E *Chair:*

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

Plaza Ballroom F *Chair:*

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.

Session 4D: Ethics and NLP

Chair:

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 $138×$*overoneofthestate − of − the − artsecurenetworksonthePennTreebankdataset.*

Session 4E: Phonology, Morphology and Word Segmentation

Chair:

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 O(n^2)$(wherenistheinputlength)orO(nm)(wheremisth$
$Corasickalgorithm.Weintroduceadditionallinkagesontopofthetriebuiltfromthevocabulary,allowing$
$tokenization(splittingthetextintowords)andourlinear-timeWordPiecemethodintoasinglepass.Expe$

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.

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

Chair:

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.

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.

5

Main Conference: Monday, November 8

Overview

	Session 5					
9:00 – 10:30	Question Answering 1 *Plaza Ballroom A & B*	Generation 2 *Plaza Ballroom D & E*	Dialogue and Interactive Systems 3 *Plaza Ballroom F*	Sentiment Analysis, Stylistic Analysis, and Argument Mining 2	Resources and Evaluation 2	Interpretability and Analysis of Models for NLP 2
11:00 – 12:00	**INVITED TALK 2 (live) by Evelina Fedorenko**					*Plaza Ballroom A, B, & C*
	Session 6					
2:45 – 4:15	Machine Learning for NLP 4 *Plaza Ballroom A & B*	Summarization 2 *Plaza Ballroom D & E*	Machine Translation and Multilinguality 2 *Plaza Ballroom F*	Speech, Vision, Robotics, Multimodal Grounding 2	Sentiment Analysis, Stylistic Analysis, and Argument Mining 3	Semantics 2
	Session 7					
4:45 – 6:15	Machine Translation and Multilinguality 3 *Plaza Ballroom A & B*	Question Answering 2 *Plaza Ballroom D & E*	Dialogue and Interactive Systems 4 *Plaza Ballroom F*	Resources and Evaluation 3	Efficient Methods for NLP 2	Semantics 3

Keynote Address: Evelina Fedorenko

The language system in the human brain.

Monday, November 8, 2021, 11:00–12:00am

Plaza Ballroom A, B, & C

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*	Contextualize Knowledge Bases with Transformer for End-to-end Task-Oriented Dialogue Systems *Gou et al.*	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.*	Efficient Dialogue Complementary Policy Learning via Deep Q-network Policy and Episodic Memory Policy *Zhao 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*	CRFR: Improving Conversational Recommender Systems via Flexible Fragments Reasoning on Knowledge Graphs *Zhou et al.*	Joint Multi-modal 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*	DuRecDial 2.0: A Bilingual Parallel Corpus for Conversational Recommendation *Liu et al.*	Solving Aspect Category Sentiment Analysis as a Text Generation Task *Liu et al.*	We Need to Talk About train-dev-test Splits *Goot*	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*	End-to-End Learning of Flowchart Grounded Task-Oriented Dialogs *Raghu, Agarwal, and Joshi*	Semantics-Preserved Data Augmentation for Aspect-Based Sentiment Analysis *Hsu et al.*	PhoMT: A High-Quality and Large-Scale Benchmark Dataset for Vietnamese-English Machine Translation *Doan et al.*	Sociolectal Analysis of Pretrained Language Models *Zhang et al.*
10:10	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*	End-to-End Learning of Flowchart Grounded Task-Oriented Dialogs *Raghu, Agarwal, and Joshi*	Semantics-Preserved Data Augmentation for Aspect-Based Sentiment Analysis *Hsu et al.*	PhoMT: A High-Quality and Large-Scale Benchmark Dataset for Vietnamese-English Machine Translation *Doan et al.*	Sociolectal Analysis of Pretrained Language Models *Zhang et al.*
10:20	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*	End-to-End Learning of Flowchart Grounded Task-Oriented Dialogs *Raghu, Agarwal, and Joshi*	Semantics-Preserved Data Augmentation for Aspect-Based Sentiment Analysis *Hsu et al.*	PhoMT: A High-Quality and Large-Scale Benchmark Dataset for Vietnamese-English Machine Translation *Doan et al.*	Sociolectal Analysis of Pretrained Language Models *Zhang et al.*

Parallel Session 5

Session 5A: Question Answering 1

Plaza Ballroom A & B *Chair:*

Improving Unsupervised Question Answering via Summarization-Informed Question Generation

Chenyang Lyu et al. 9:00–9:15

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. 9:15–9:30

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. 9:30–9:45

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. 9:45–10:00

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

Plaza Ballroom D & E *Chair:*

DiscoDVT: Generating Long Text with Discourse-Aware Discrete Variational Transformer

Haozhe Ji and Minlie Huang 9:00–9:15

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. 9:15–9:30

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 9:30–9:45

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 9:45–10:00

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

Plaza Ballroom F *Chair:*

Contextualize Knowledge Bases with Transformer for End-to-end Task-Oriented Dialogue Systems

Yanjie Gou et al. 9:15–9:30

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. 9:30–9:45

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. 9:45–10:00

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, 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.

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

Chair:

Dimensional Emotion Detection from Categorical Emotion

Sungjoon Park et al. 9:00–9:15

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 9:15–9:30

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. 9:30–9:45

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. 9:45–10:00

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

Chair:

CSDS: A Fine-Grained Chinese Dataset for Customer Service Dialogue Summarization

Haitao Lin et al. 9:00–9:15

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. 9:15–9:30

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. 9:30–9:45

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.

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

Chair:

Multi-granularity Textual Adversarial Attack with Behavior Cloning

Yangyi Chen, Jin Su, and Wei Wei 9:00–9:15

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 9:15–9:30

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. 9:30–9:45

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. 9:45–10:00

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.

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 Multitask 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.*	Generating Datasets with Pretrained Language Models *Schick and Schütze*	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.*	Continuous Entailment Patterns for Lexical Inference in Context *Schmitt and Schütze*
4:00	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.*	Continuous Entailment Patterns for Lexical Inference in Context *Schmitt and Schütze*

Parallel Session 6

Session 6A: Machine Learning for NLP 4

Plaza Ballroom A & B *Chair:*

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

Plaza Ballroom D & E *Chair:*

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.

Session 6C: Machine Translation and Multilinguality 2

Plaza Ballroom F *Chair:*

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

Chair:

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

Chair:

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

Chair:

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.

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.

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*	
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*	4:45
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.*	5:00
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.*	5:15
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.*	5:30

	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*
5:45	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*
6:00	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*

Parallel Session 7

Session 7A: Machine Translation and Multilinguality 3

Plaza Ballroom A & B *Chair:*

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

Plaza Ballroom D & E *Chair:*

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/.

Session 7C: Dialogue and Interactive Systems 4

Plaza Ballroom F Chair:

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

Chair:

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

Chair:

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})$\$inthelengthoftextandthenumberofpotentiallinksis\$O(n^{4})\$, var$

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

Chair:

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 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).

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.

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

Session 8

11:00 – 12:30	Machine Learning for NLP 5 *Plaza Ballroom A & B*	Question Answering 3 *Plaza Ballroom D & E*	Information Extraction 4 *Plaza Ballroom F*	NLP Applications 2	Speech, Vision, Robotics, Multimodal Grounding 3	Semantics 4

Session 9

2:30 – 4:00	Machine Translation and Multilinguality 4 *Plaza Ballroom A & B*	Interpretability and Analysis of Models for NLP 3 *Plaza Ballroom D & E*	Information Extraction 5 *Plaza Ballroom F*	Resources and Evaluation 4	Syntax	Efficient Methods for NLP 3

4:30 – 5:30	**INVITED TALK 3 (recorded) by Steven Bird**	*Plaza Ballroom A, B, & C*
5:30 – 6:00	**Closing**	

Keynote Address: Steven Bird

LT4All!? Rethinking the Agenda

Tuesday, November 9, 2021, 16:30–17:30am

Plaza Ballroom A, B, & C

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.*	QA-Align: Representing Cross-Text Content Overlap by Aligning Question-Answer Propositions *Brook Weiss 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.*	PICARD: Parsing Incrementally for Constrained Auto-Regressive Decoding from Language Models *Scholak, Schucher, and Bahdanau*	11:30
Knowledge Graph Representation Learning using Ordinary Differential Equations *Nayyeri et al.*	What's in a Name? Answer Equivalence For Open-Domain Question Answering *Si, Zhao, and Boyd-Graber*	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*	Exploiting Twitter as Source of Large Corpora of Weakly Similar Pairs for Semantic Sentence Embeddings *Di Giovanni and Brambilla*	11:45
KnowMAN: Weakly Supervised Multinomial Adversarial Networks *März et al.*	Evaluation Paradigms in Question Answering *Rodriguez 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.*	Guilt by Association: Emotion Intensities in Lexical Representations *Raji and Melo*	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	KnowMAN: Weakly Supervised Multinomial Adversarial Networks *März et al.*	Evaluation Paradigms in Question Answering *Rodriguez 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.*	Guilt by Association: Emotion Intensities in Lexical Representations *Raji and Melo*
12:20	KnowMAN: Weakly Supervised Multinomial Adversarial Networks *März et al.*	Evaluation Paradigms in Question Answering *Rodriguez 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.*	Guilt by Association: Emotion Intensities in Lexical Representations *Raji and Melo*

Parallel Session 8

Session 8A: Machine Learning for NLP 5

Plaza Ballroom A & B *Chair:*

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

Plaza Ballroom D & E *Chair:*

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.

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

Plaza Ballroom F *Chair:*

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

Chair:

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

Chair:

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.

Session 8F: Semantics 4

Chair:

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.

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.

Session 9 Overview – Tuesday, November 9, 2021

	Track A	**Track B**	**Track C**
	Machine Translation and Multilinguality 4 Plaza Ballroom A & B	*Interpretability and Analysis of Models for NLP 3* Plaza Ballroom D & E	*Information Extraction 5* Plaza Ballroom F
2:30	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.*
2:45	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*
3:00	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.*	Reducing Discontinuous to Continuous Parsing with Pointer Network Reordering *Fernández-González and Gómez-Rodríguez*	Block Pruning For Faster Transformers *Lagunas et al.*
3:15	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.*	A New Representation for Span-based CCG Parsing *Kato and Matsubara*	Finetuning Pretrained Transformers into RNNs *Kasai et al.*

Parallel Session 9

Session 9A: Machine Translation and Multilinguality 4

Plaza Ballroom A & B *Chair:*

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.

Session 9B: Interpretability and Analysis of Models for NLP 3

Plaza Ballroom D & E *Chair:*

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

Plaza Ballroom F *Chair:*

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

Chair:

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

Chair:

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}), $whichisoftengreaterthanthemeanhittingtimeofac$
$K^{2}N)time.Tothebestofourknowledge,noalgorithmhasbeengivenforsamplingspanningtreeswithoutr$

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

Chair:

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 tradeoff 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.

7

Workshops

Wednesday–Thursday

Sixth Conference on Machine Translation p.??
The 25th Conference on Computational Natural Language Learning (CoNLL) p.162
8th International Workshop on Argument Mining p.165
2nd Workshop on Computational Approaches to Discourse (CODI) p.166

Wednesday

3rd Workshop on NLP for Conversational AI (NLP4ConvAI) p.169
Novel Ideas in Learning-to-Learn through Interaction p.??
2nd Workshop on Evaluation and Comparison of NLP Systems p.??
CI+NLP: First Workshop on Causal Inference in NLP p.171
The Second Workshop on Insights from Negative Results in NLP p.172
The 3rd Workshop on Machine Reading for Question Answering p.173
The Natural Legal Language Processing Workshop 2021 (NLLP) p.176
SustaiNLP 2021: The Second Workshop on Simple and Efficient Natural Language Processing p.177
Fourth Workshop on Fact Extraction and VERification (FEVER) p.178
Third Workshop on New Frontiers in Summarization p.179

Thursday

3rd Workshop on Economics and Natural Language Processing (ECONLP) p.181
LAW-DMR: The 15th Linguistic Annotation and 3rd Designing Meaning Representations Joint Workshop p.??
7th Workshop on Noisy User-generated Text (W-NUT 2021) p.182
The 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL) p.184
The 1st Workshop on Multi-lingual Representation Learning (MRL) p.??
The Fifth Widening NLP Workshop (WiNLP 2021) p.??
Workshop on Evaluations and Assessments of Neural Conversation Systems p.185
Fourth Workshop on Computational Models of Reference, Anaphora and Coreference p.186
BlackboxNLP: Analyzing and interpreting neural networks for NLP p.188

Workshop 22: The 25th Conference on Computational Natural Language Learning (CoNLL)

Organizers: *Omri Abend et al.*

Venue: Governor's Square 12

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*

Venue: Governor's Square 12

- Argument Mining on Twitter: A Case Study on the Planned Parenthood Debate
  Muhammad Mahad Afzal Bhatti, Ahsan Suheer Ahmad, and Joonsuk Park
- Multi-task and Multi-corpora Training Strategies to Enhance Argumentative Sentence Linking Performance
  Jan Wira Gotama Putra, Simone Teufel, and Takenobu Tokunaga
- Explainable Unsupervised Argument Similarity Rating with Abstract Meaning Representation and Conclusion Generation
  Juri Opitz et al.
- 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.
- 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.
- 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.
- 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
- 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

Workshop 16: 2nd Workshop on Computational Approaches to Discourse (CODI)

Organizers: *Chloé Braud et al.*

Venue: Governor's Square 12

Wednesday, November 10, 2021

9:00–12:00 **CODI-CRAC Shared Task**

9:05–9:30 **Welcome**

9:05–9:30 **The CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis Resolution in Dialogue (Sopan Khosla, Juntao Yu, Ramesh Manuvinakurike, Vincent Ng, Massimo Poesio, Michael Strube and Carolyn Rosé)**

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.*

Venue: Governor's Square 12

- 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 6: CI+NLP: First Workshop on Causal Inference in NLP

Organizers: *Jacob Eisenstein et al.*

Venue: Director's Row J

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.*

Venue: Governors Square 11

- 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.*

Venue: Governors Square 11

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.*

Venue: Governors Square 11

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.

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.*

Venue: Governors Square 11

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.*

Venue: Governors Square 11

Wednesday, November 10, 2021

0900–0905 Opening

0905–1020 Keynote Talks

1020–1100 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.

11:00–11:15 **Break**

1115–1230 Keynote Talks

12:30–1:30 **Lunch**

1330–1445 Keynote and 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

1445–1615 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.

1615–1730 Keynote Talks

5:30–5:45 **Closing Remarks**

Workshop 21: Third Workshop on New Frontiers in Summarization (NewSum)

Organizers: *Lu Wang et al.*

Venue: Governors Square 11

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*

Venue: Governors Square 17

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 7: 7th Workshop on Noisy User-generated Text (W-NUT 2021)

Organizers: *Wei Xu et al.*

Venue: Director's Row I

- 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.*

Venue: Director’s Row J

- 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 17: Workshop on Evaluations and Assessments of Neural Conversation Systems

Organizers: *Wei Wei et al.*

Venue: Director's Row H

- 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.*

Venue: Director's Row H

Wednesday, November 10, 2021

9:00–10:00 **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-TalkingRobots System for the CODI-CRAC 2021 Shared-Task**

10:00–10:30 **Coffee Break**

10:30–12:00 **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**

9:05–10:00 **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.

11:00–12:00 **Invited Talk**

11:00–12:00 **Modeling informational relations across multiple texts**

12:00–1:00 **Lunch Break**

1:00–2:30 **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**

2:45–2:55 **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**

3:45–4:15 **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**

4:45–5:50 **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.*

Venue: Director's Row H

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
	Poster Sessions
	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

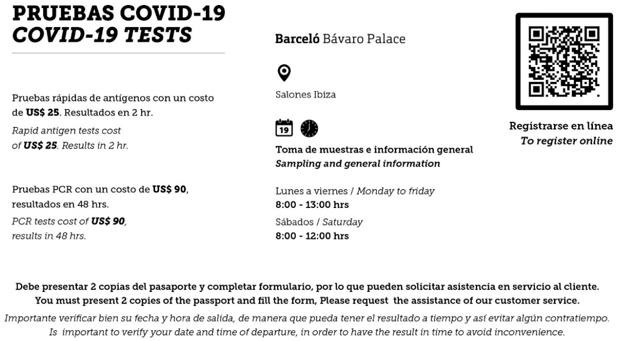
Revised by ACL Executive Committee, July 15, 2018

The policy is also available from ACL's main page.

9

Local Guide

COVID-19 Tests



PRUEBAS COVID-19
COVID-19 TESTS

Pruebas rápidas de antígenos con un costo de **US$ 25**. Resultados en 2 hr.

*Rapid antigen tests cost of **US$ 25**. Results in 2 hr.*

Pruebas PCR con un costo de **US$ 90**, resultados en 48 hrs.

*PCR tests cost of **US$ 90**, results in 48 hrs.*

Barceló Bávaro Palace

Salones Ibiza

Toma de muestras e información general
Sampling and general information

Lunes a viernes / *Monday to friday*
8:00 - 13:00 hrs
Sábados / *Saturday*
8:00 - 12:00 hrs

Registrarse en línea
To register online

Debe presentar 2 copias del pasaporte y completar formulario, por lo que pueden solicitar asistencia en servicio al cliente.
You must present 2 copies of the passport and fill the form, Please request the assistance of our customer service.
Importante verificar bien su fecha y hora de salida, de manera que pueda tener el resultado a tiempo y así evitar algún contratiempo.
Is important to verify your date and time of departure, in order to have the result in time to avoid inconvenience.

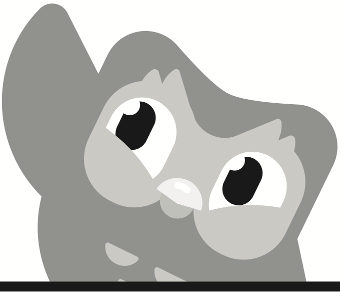
Transportation and Tours:

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.



EMNLP 2021 gratefully acknowledges the following sponsors for their support:

Diamond

Bloomberg
Engineering

FACEBOOK AI

Google Research

Platinum

amazon | science

DeepMind

RESEARCH

Megagon Labs

Gold

grammarly

Microsoft

Silver

NAVER

NAVER LABS
Europe

Bronze

Adobe

Babelscape

human language technology
center of excellence

LegalForce

Supporter

servicenow

Diversity and Inclusion: Champion

amazon | science

DeepMind

Google Research

Microsoft

Diversity and Inclusion: Ally

Morgan Stanley

Diversity and Inclusion: Contributor

Adobe

IBM